SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 1–32, January 1997

001

ON THE MELTING OF ICE BALLS[*]

MIGUEL A. HERRERO[†] AND JUAN J. L. VELÁZQUEZ[†]

Abstract. We consider here the problem of describing the melting of an ice ball surrounded by water. The corresponding mathematical model consists of the Stefan problem with radial symmetry. We obtain asymptotic expansions for the radius of the melting ball which turn out to be of a different nature according to the cases $N \geq 3$ and $N = 2$, N being the space dimension. The methods employed combine matched asymptotic expansion techniques, a priori estimates, and topological results.

Key words. Stefan problem, asymptotic behavior, matched asymptotic expansions, a priori estimates

AMS subject classifications. 35R55, 35B40, 35C20

PII. S0036141095282152

1. Introduction.

This paper is concerned with radial solutions of the following Stefan problem. To find functions $\theta(r,t)$ and $R(t)$ such that

$$\theta_t = \theta_{rr} + \left(\frac{N-1}{r}\right)\theta_r \quad \text{for } r > R(t) \quad \text{and} \quad t > 0 \tag{1.1a}$$

$$\theta(r,0) = \theta_0(r) \quad \text{for } r > R(0), \tag{1.1b}$$

$$\theta(r,t) = 0 \quad \text{if } r \leq R(t) \quad \text{and} \quad t > 0, \tag{1.1c}$$

$$\theta_r(R(t),t) = -\dot{R}(t) \quad \text{if } t > 0. \tag{1.1d}$$

Here $r = |x|$, $x \in \mathbb{R}^N$, and $N \geq 2$. As it stands, (1.1) is a model for describing the melting of a ball of ice surrounded by water. $\theta(r,t)$ denotes the temperature of the medium, which is assumed to be zero at the ice phase. We do not require $\theta_0(r)$ to be positive for every $r > R(0)$, so that the existence of regions where water is initially undercooled is not ruled out.

In view of classical results, one expects that under fairly general circumstances (for instance, if $\theta_0(r)$ is nonnegative or if undercooling does not affect the dynamics of the problem much) the ice ball will entirely melt at some finite time $t = T < \infty$. We shall address the following here.

Question. What is the speed at which ice balls collapse? In other words, what is the asymptotic behavior of $R(t)$ as $t \uparrow T$?

We shall show in what follows that there is a countable family of possible behaviors for $R(t)$ as the melting time $t = T$ is approached. To describe our results, it will be convenient to consider separately the cases $N = 2$ and $N \geq 3$. In the bidimensional situation we prove the following theorem.

[*] Received by the editors February 27, 1995; accepted for publication (in revised form) October 4, 1995. This research was partially supported by DGICYT grant PB93-0438 and EEC contract CHRX-CT93-0413.

http://www.siam.org/journals/sima/28-1/28215.html

[†] Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense, 28040 Madrid, Spain (herrero@sunma1.mat.ucm.es, velazquez@sunma1.mat.ucm.es).

Theorem 1.1. *Assume that $N = 2$. For any $T > 0$, there exist solutions of* (1.1) *such that the corresponding interfaces behave as $t \uparrow T$ in one of the following ways:*

$$R(t) = B(T-t)^{\frac{1}{2}} e^{-\frac{\sqrt{2}}{2}|\log(T-t)|^{\frac{1}{2}}} |\log(T-t)|^{\frac{1}{4\sqrt{\log(T-t)}}-\frac{1}{4}} \cdot (1+o(1)), \tag{1.2}$$

where B is a fixed positive constant, or

$$R(t) = C(T-t)^{\frac{l}{2}} |\log(T-t)|^{-\frac{l}{2(l-1)}} (1+o(1)), \tag{1.3}$$

where C is an arbitrary positive constant and l is any integer such that $l \geq 2$.

Concerning the case of higher dimensions, we obtain the following theorem.

Theorem 1.2. *Assume that $N \geq 3$. For any $T > 0$, there exist solutions of* (1.1) *such that the corresponding interfaces behave as $t \uparrow T$ in one of the following ways:*

$$R(t) = B_N (T-t)^{\frac{1}{2}} |\log(T-t)|^{-\frac{1}{N-2}} (1+o(1)), \tag{1.4}$$

where B_N is a fixed positive constant depending on the dimension N,

$$\textit{or} \qquad R(t) = C(T-t)^{\frac{l}{2}} (1+o(1)), \tag{1.5}$$

where C is an arbitrary positive constant and l is any integer number such that $l \geq 2$.

Let us remark briefly on Theorems 1.1 and 1.2. To begin with, we do not preclude here the existence of other possible types of shrinking spheres besides those described in (1.2)–(1.5), although it seems very unlikely in view of the arguments leading to the proofs of these results. As a matter of fact, we expect (1.2) and (1.4) to provide the generic asymptotics for the case of the classical, not undercooled, Stefan problem. However, no proof of such a statement is given here. It will be apparent from the proofs that (1.3) and (1.5) correspond to problems with small undercooling, i.e., problems where temperature changes sign somewhere for any $t < T$.

We next observe that (1.2)–(1.5) imply that the ice radii $R(t)$ are such that

$$R(t) \ll (T-t)^{\frac{1}{2}} \quad \text{as } t \uparrow T,$$

and the contracting rates are therefore faster than those corresponding to the natural scales of the problem under consideration. In particular, the solutions obtained are not self-similar.

It is worth pointing out that our approach here allows us to obtain further information on the structure of the solutions involved. For instance, asymptotic expansions for the predicted water temperature near the melting ice ball can be obtained as $t \uparrow T$. The corresponding result reads as follows.

Theorem 1.3. *Assume first that $N = 2$. Then the solutions referred to in Theorem* 1.1 *are such that the following expansions hold:*

If (1.2) *occurs, then*

$$\theta(x,T) = D\ e^{-2|\log|x||^{\frac{1}{2}}} |2\log|x||^{\frac{1}{2\sqrt{2|\log|x||}}} (1+o(1)) \quad \textit{as } x \downarrow 0, \tag{1.6}$$

where D is a fixed positive constant.

If (1.3) *occurs, then*

$$\theta(x,T) = D_1 |x|^{2l-2} (|\log|x||)^{-\frac{1}{l-1}} (1+o(1)) \quad \textit{as } x \downarrow 0 \tag{1.7}$$

for some positive constant D_1.

Now suppose that $N \geq 3$. *Then the solutions referred to in Theorem* 1.2 *are such that the following expansions hold:*

If (1.4) *occurs, then*

$$\theta(x,T) = K_N(\log|x|)^{-\frac{2}{N-2}}(1+o(1)) \quad \textit{as } x \downarrow 0, \tag{1.8}$$

where K_N *is a fixed positive constant depending on the dimension* N.

If (1.5) *occurs, then*

$$\theta(x,T) = K_1|x|^{2l-2}(1+o(1)) \quad \textit{as } x \downarrow 0 \tag{1.9}$$

for some positive constant K_1.

Concerning previous related work, it is well known that for $N = 1$ the asymptotic shape of the vanishing ice phase is a space-time wedge which has its tip at $t = T$. We refer for such a case to the paper [AK], where disappearance of one of the phases in a one-dimensional, two-phase Stefan problem is studied by means of functional analysis methods. A different approach, based on matched asymptotics expansion techniques has been developed in [RSP] and [SW].

For instance, in [SW] a formal analysis of (1.1) with (1.1d) replaced by

$$\theta_r(R(t),t) = -\Lambda\dot{R}(t) \quad \text{for } t > 0 \tag{1.10}$$

is performed in the limit $\Lambda \to \infty$. In particular a boundary layer is then identified where an expansion similar to (1.2) and (1.4) takes place. The reader is referred to [DH], [HD1], [HD2], and [S] for related work, as well as to [R] and [M] for a general outline of results concerning Stefan problems.

We conclude this introduction by describing the plan of the paper. Some preliminary material is gathered in section 2. Section 3 is then devoted to deriving the results in Theorems 1.1–1.3 by means of matched asymptotic expansion techniques in a way which we believe to be conceptually simpler than the study done in [RSP], [SW] for (1.1a)–(1.1c) and (1.10). Besides its heuristic interest, this formal method detects a number of previously unnoticed patterns and provides the basic lines along which a rigorous proof is subsequently implemented. The arguments in section 3 are made rigorous in sections 4 to 7. To be precise, the sought-for solutions are obtained by means of a topological fixed point argument. This is a classical approach in the literature on partial differential equations (PDEs) which, to mention but a few examples, has been used recently to analyze singularity patterns arising in parabolic equations in the works [B1], [B2], [AV], and [HV3], among others. The basic aspects of our topological method are presented in section 4.

Sections 5 and 6 are then devoted to providing the various estimates required to yield (1.2) in Theorem 1.1 and (1.6) in Theorem 1.3. Once this has been achieved, we conclude by sketching in section 7 those modifications required to obtain (1.3)–(1.5) as well as (1.7)–(1.9).

2. Preliminaries.

Let $(\theta(r,t),\, R(t))$ be a solution of (1.1). It will be convenient for our purposes to introduce a new variable $u(r,t)$ given by

$$u(r,t) = \int_{R(t)}^{r} \xi^{1-N}d\xi \int_{R(t)}^{\xi} s^{N-1}(\theta(s,t)+1)\,ds \quad \text{if } r > R(t), \tag{2.1a}$$

$$u(r,t) = 0 \quad \text{if } r \leq R(t). \tag{2.1b}$$

We then readily see that $u(r,t)$ satisfies

$$u_t = u_{rr} + \left(\frac{N-1}{r}\right) u_r - H(u) \quad \text{for } x \in \mathbb{R}^N, \quad t > 0, \tag{2.2a}$$

where $H(u)$ is the standard Heaviside function; i.e., $H(u) = 1$ whenever $u > 0$ and $H(u) = 0$ otherwise. Equation (2.2a) is to be complemented with the initial condition

$$u(r,0) = \int_{R(0)}^{r} \xi^{1-N} d\xi \int_{R(0)}^{\xi} s^{N-1}(\theta(s,0)+1)\, ds. \tag{2.2b}$$

For ease of notation, we shall often use the symbols Δ and ∇ instead of their radial counterparts when dealing with (2.2a) and related equations. Further, we introduce self-similar variables as follows:

$$u(r,t) = (T-t)\Phi(y,\tau), \quad y = r(T-t)^{-\frac{1}{2}}, \quad \tau = -\log(T-t); \tag{2.3}$$

we define a rescaled free boundary $\varepsilon(\tau)$ given by

$$\varepsilon(\tau) = R(t)e^{\frac{\tau}{2}} \equiv R(t)(T-t)^{-\frac{1}{2}}. \tag{2.4}$$

It is then readily seen that Φ satisfies the following equation:

$$\begin{aligned} \Phi_\tau &= \Phi_{yy} + \left(\frac{N-1}{y} - \frac{y}{2}\right)\Phi_y - H(\Phi) \\ &\equiv \Delta\Phi - \frac{1}{2}y\nabla\Phi + \Phi - (1-\chi_\varepsilon) \\ &\equiv A\Phi - (1-\chi_\varepsilon), \end{aligned} \tag{2.5}$$

where $\chi_\varepsilon(y) = 1$ for $y < \varepsilon(\tau)$ and $\chi_\varepsilon(y) = 0$ otherwise. The linear operator A will play a key role in our approach. Consider the weighted space

$$L^2_{w,r}(\mathbb{R}^+) = \left\{ f \in L^2_{\text{loc}}(\mathbb{R}^+) : \|f\|^2 = \int_0^\infty y^{N-1}|f(y)|^2 e^{\frac{-y^2}{4}}\, dy < \infty \right\}.$$

Clearly $L^2_{w,r}(\mathbb{R}^+)$ is a Hilbert space when endowed with the norm

$$\|f\|^2 = \langle f, f\rangle = \int_0^\infty y^{N-1}|f(y)|^2 e^{\frac{-y^2}{4}}\, dy,$$

where we have used the symbol $\langle\ ,\ \rangle$ to denote the corresponding scalar product. For any positive integer k, one may then define the Hilbert spaces $H^k_{w,r}(\mathbb{R}^+)$ in a straightforward way. By classical spectral theory, one then has that the radial operator A in (2.5) is self-adjoint in $L^2_{w,r}(\mathbb{R}^+)$ with domain $D(A) = H^2_{w,r}(\mathbb{R}^+)$. Furthermore, the eigenvalues of A consist of the sequence

$$\lambda_k = 1-k, \quad k = 0,1,2,\ldots. \tag{2.6a}$$

The corresponding eigenfunctions can be written in the form

$$\varphi_k(y) = \begin{cases} c_k L_k\left(\frac{y^2}{4}\right) & \text{if } N = 2, \quad k = 0,1,2,\ldots, \\ c_{k,N} L_k^{\frac{N-2}{N}}\left(\frac{y^2}{4}\right) & \text{if } N \geq 3, \quad k = 0,1,2,\ldots, \end{cases} \tag{2.6b}$$

where $L_k(x)$ (resp. $L_k^{(N-2)/N}(x)$) denotes the standard kth Laguerre polynomial (resp. the modified $L_k^{\alpha}(x)$ Laguerre polynomial with $\alpha = \frac{N-2}{N}$), cf., for instance, [L] and [MF] for a review of properties of such functions. The normalization constants c_k and $c_{k,N}$ in (2.6b) are selected so that

$$\|\varphi_k\| = 1 \quad \text{for any } k. \tag{2.6c}$$

By classical results (cf., for instance, [MF]), we readily see that

$$c_{k,N}^2 = \frac{\Gamma(\frac{N}{2})\Gamma(k+1)}{2^N \left(\Gamma\left(\frac{N-2}{2}+k+1\right)\right)^2 \pi^{\frac{N}{2}}}, \tag{2.7a}$$

$$c_k^2 = \frac{1}{4\pi\Gamma(k+1)}. \tag{2.7b}$$

The following a priori bound on solutions of (2.5) is important for our purposes:

$$\Phi(y,\tau) \le C(y^2+1) \quad \text{for some } C>0 \text{ and any } y>0, \tau>0. \tag{2.8}$$

Estimate (2.8) can be obtained, for instance, from the Bernstein-type bound

$$|\nabla\Phi(y,\tau)| \le C \quad \text{for any } y \text{ and any } \tau > 0,$$

which holds for solutions of (2.5) under rather loose assumptions on their initial values (cf., for instance, [HV1] for a related result). Arguing as in [HV2], we may deduce from (2.8) the following convergence result:

$$\begin{aligned} &\Phi(y,t) \to \frac{y^2}{4} \quad \text{as } t \to \infty, \\ &\text{uniformly on sets } y \le M < \infty. \end{aligned} \tag{2.9}$$

Since (2.9) plays an important role in what follows, we shall briefly sketch here the main ideas in its proof and refer to [HV2] for details. To begin with, an energy argument like that in [GK] shows that

$$\Phi(y,\tau) \to \Phi^*(y) \quad \tau \to \infty$$

uniformly on compact sets of $|y|$, where Φ^* is a stationary solution of (2.5). Arguing as in Lemma 4.3 of [HV2], we see that either $\lim_{\tau\to\infty} \varepsilon(\tau) = 0$ or $\lim_{\tau\to\infty} \varepsilon(\tau) = 1$. The second case would allow for a possible stationary solution $\Phi^*(y) = 1$, which would in turn yield that $\Phi(0,\tau) > 0$ for $\tau \gg 1$. This in particular implies that the ice ball has already disappeared for some time $t < T$, which is a contradiction. On the other hand, the case $\lim_{\tau\to\infty} \varepsilon(r) = 0$ gives rise to two possible stationary solutions satisfying (2.8), namely

$$\Phi^*(y) = 0, \qquad \Phi^*(y) = \frac{y^2}{4}.$$

To rule out the first possibility, we argue by contradiction as follows. Assume that $\lim_{\tau\to\infty} \Phi(y,r) = 0$ uniformly on sets $|y| \le R < \infty$. Then for fixed $A > 0$ and $\varepsilon > 0$ we may select $\tau \gg 1$ so that $\Phi(y,\tau) \le \varepsilon$ for $\tau \ge \tau_0$ and $|y| \le A$. A quick glance at equation (2.5) reveals then that $\Phi(y,\tau)$ is at most of order $O(\varepsilon e^{\tau-\tau_0})$ for $\tau > \tau_0$

at distances $y \sim Ae^{(\tau-\tau_0)/2}$. As a matter of fact, this estimate is readily suggested by dropping the absorption term $H(\Phi)$ in (2.5) and then checking how bounds on initial values propagate along characteristics for the resulting equation. In terms of the variable $u(x,t)$, one is thus led to a bound of the type

$$u(x,t) \le \varepsilon x^2 \quad \text{for} \quad x \le \delta, \quad t_0 < t < T, \tag{2.10}$$

where $\delta = \delta(\varepsilon) > 0$ is a small (but fixed) positive number and t_0 is close enough to T. On the other hand, our assumption $\Phi^*(y) = 0$ carries into

$$u(x,t_0) \le \varepsilon(T-t_0) \quad \text{for } x \le A. \tag{2.11}$$

From (2.10) and (2.11), a barrier argument as the one in [EK] or [HV1] yields that $u(x,T) = 0$ for some $x > 0$, thus contradicting the assumption that the ice ball collapses exactly at $t = T$. This concludes the proof.

3. The formal argument.

This section is devoted to showing how to obtain the asymptotic results in Theorems 1.1 and 1.2 by means of formal perturbative methods. While the approach to be described is a nonrigorous one, it is in our opinion the crux of this work. The reason for this statement is that these heuristic methods not only provide deep insight into what to expect but also mark the path along which a rigorous argument can be implemented. This last task will be postponed until sections 4–7.

For definiteness, we shall consider first the case $N = 2$ and remark then about the differences which arise for $N \ge 3$. Our starting point is the convergence result (2.9). Bearing it in mind, we set

$$\psi(y,\tau) = \Phi(y,\tau) - \frac{y^2}{4} \tag{3.1}$$

so that the function $\psi(y,\tau)$ satisfies

$$\psi_\tau = A\psi + \chi_{\varepsilon(\tau)}. \tag{3.2}$$

We now introduce the following ansatz concerning the effect of the term $\chi_{\varepsilon(\tau)}$ in (3.2).

Assumption 3.1. For $|y| \gg \varepsilon(\tau)$ and $\tau \gg 1$, we may replace (3.2) by

$$\psi_\tau = A\psi + \gamma\varepsilon(\tau)^2\delta(y), \tag{3.3}$$

where constant γ is uniquely determined by imposing that

$$\int_{\mathbb{R}^2} \chi_{\varepsilon(\tau)} dy = \gamma\varepsilon(\tau)^2 \int_{\mathbb{R}^2} \delta(y) dy \quad \text{(i.e., } \gamma = \pi). \tag{3.4}$$

We next proceed to derive (1.2) in Theorem 1.1. To this end, we set

$$\Psi(y,\tau) = \sum_{k=0}^{\infty} a_k(\tau)\varphi_k(y) \equiv a_0(\tau)\varphi_0(y) + a_1(\tau)\varphi_1(y) + Q(y,\tau). \tag{3.5}$$

The first two Fourier coefficients would then satisfy

$$\dot{a}_0 = a_0 + \gamma\varepsilon(\tau)^2\langle\varphi_0, \delta(y)\rangle \quad \text{for } \tau \gg 1, \tag{3.6a}$$

$$\dot{a}_1 = \gamma\varepsilon(\tau)^2\langle\varphi_1, \delta(y)\rangle \quad \text{for } \tau \gg 1, \tag{3.6b}$$

whereas the remainder term $Q(y,\tau)$ is such that

$$Q_\tau = AQ + \gamma\varepsilon(\tau)^2 \left(\delta(y) - \sum_{k=0}^{1} \langle\varphi_k, \delta(y)\rangle\varphi_k \right), \tag{3.6c}$$

$$\langle Q, \varphi_k \rangle = 0 \quad \text{for } k = 0, 1. \tag{3.6d}$$

We now introduce the following assumption.

Assumption 3.2. The leading term in (3.5) as $\tau \gg 1$ is $a_1(\tau)\varphi_1(y)$; i.e., evolution in time of $\psi(y, \tau)$ is driven by the eigenfunction corresponding to zero eigenvalue. Moreover, one then expects

$$|\dot{\varepsilon}(\tau)| \ll \varepsilon(\tau) \quad \text{as } \tau \to \infty, \tag{3.7a}$$

$$Q(y, \tau) \sim \gamma \varepsilon(\tau)^2 F(y) \quad \text{as } \tau \to \infty \tag{3.7b}$$

for a suitable function $F(y)$. It then turns out that $F(y)$ satisfies

$$AF + \left(\delta(y) - \sum_{k=0}^{1} \langle \varphi_k, \delta(y) \rangle \varphi_k \right) = 0, \tag{3.8a}$$

$$\langle F, \varphi_k \rangle = 0 \quad \text{for } k = 0, 1. \tag{3.8b}$$

We may now integrate (3.8a) and (3.8b) to obtain

$$F(y) = -\frac{1}{2\pi} \log y + B + O(y^2 |\log y|) \quad \text{for } \varepsilon(\tau) \ll y \le 1, \tag{3.9}$$

where constant B is detemined by the orthogonality conditions (3.8b). On the other hand, since we expect $\lim_{\tau \to \infty} a_k(\tau) = 0$ for $k = 0, 1$, we obtain from (3.6) that

$$a_k(\tau) \sim -\gamma \varphi_k(0) \int_\tau^\infty \varepsilon(s)^2 e^{(1-k)(\tau - s)} ds \quad \text{for } \tau \gg 1, \quad k = 0, 1. \tag{3.10}$$

Putting together (3.5), (3.9), and (3.10), we arrive at

$$\Phi(y, \tau \sim \frac{y^2}{4} - \gamma \sum_{k=0}^{1} \varphi_k(0)\varphi_k(y) \int_\tau^\infty \varepsilon(s)^2 e^{(1-k)(\tau - s)} ds + \gamma \varepsilon(\tau)^2 \left(B - \frac{1}{2\pi} \log y \right) \tag{3.11}$$

whenever

$$\varepsilon(\tau) \ll y \le C \quad \text{with } C > 0 \text{ and } \tau \gg 1.$$

Formula (3.11) provides an outer expansion for $\Phi(y, \tau)$ in regions sufficiently far from the free boundary. To analyze the set where $y \sim \varepsilon(\tau)$, we change variables as follows:

$$\Phi(y, \tau) = (\varepsilon(\tau))^2 w(\xi, \tau), \quad \xi = \frac{y}{\varepsilon(\tau)}. \tag{3.12}$$

Substituting (3.12) into (2.5) readily gives

$$\varepsilon \dot{\varepsilon} w - \dot{\varepsilon} \varepsilon \xi w_\xi + \varepsilon^2 w_\tau = \Delta w - \frac{\xi^2}{2} \xi w_\xi + \varepsilon^2 w - \tilde{\chi}, \tag{3.13}$$

where now $\Delta w = w_{\xi\xi} + \frac{w_\xi}{\xi}$ and $\tilde{\chi}(\xi) = 1$ whenever $\xi > 1$ and is zero elsewhere. After comparing the order of magnitude of the different terms in (3.13), we are led to guess that as $\tau \to \infty$, $w(\xi, \tau) \sim \bar{w}(\xi)$, where $\bar{w}(\xi)$ is the solution of

$$w_{\xi\xi} + \frac{w_\xi}{\xi} = 1 \quad \text{for } \xi > 1, \tag{3.14}$$

$$w(1) = w_\xi(1) = 0;$$

i.e.,

$$\bar{w}(\xi) = \frac{\xi^2}{4} - \frac{1}{2}\log\xi - \frac{1}{4} \quad \text{for } \xi > 1. \tag{3.15}$$

In view of (3.12) and (3.15), we expect that

$$\Phi(y, \tau) \sim \varepsilon^2(\tau)\bar{w}\left(\frac{y}{\varepsilon(\tau)}\right) = \frac{y^2}{4} - \frac{\varepsilon(\tau)^2}{2}\log\left(\frac{y}{\varepsilon(\tau)}\right) - \frac{\varepsilon(\tau)^2}{4} \tag{3.16}$$

for $y \sim \varepsilon(\tau)$ and $\tau \gg 1$.

Matching the inner and outer expansions (3.16) and (3.11), we obtain as a matching condition

$$B\varepsilon(\tau)^2 - \gamma\sum_{k=0}^{1}\varphi_k(0)^2\int_\tau^\infty \varepsilon(s)^2 e^{(1-k)(\tau-s)}ds = \frac{\varepsilon(\tau)^2}{2}\log\varepsilon(\tau) - \frac{\varepsilon(\tau)^2}{4}. \tag{3.17}$$

This is the basic integral equation that determines the position of the rescaled free boundary $\varepsilon(\tau)$. Actually, we claim that (3.17) yields

$$\varepsilon(\tau) \sim Ke^{-\frac{\sqrt{2\tau}}{2}}\tau^{\frac{1}{4\sqrt{\tau}}-\frac{1}{4}} \cdot (1 + o(1)) \quad \text{as } \tau \to \infty, \tag{3.18}$$

where $K = e^{4B}$.

Taking into account (2.4), one readily checks that (3.18) gives (1.2) in Theorem 1.1. For the convenience of the reader we shall briefly sketch the way in which (3.18) can be derived from (3.17). We first observe that a dominated balance argument shows that the leading terms in (3.17) satisfy

$$\frac{\varepsilon(\tau)^2}{4}\log(\varepsilon(\tau)^2) \sim -\gamma\varphi_1(0)^2\int_\tau^\infty \varepsilon(s)^2 ds = -\frac{1}{4}\int_\tau^\infty \varepsilon(s)^2 ds. \tag{3.19}$$

Now set

$$G(\tau) = \int_\tau^\infty \varepsilon(s)^2 ds.$$

Then

$$G(\tau) = -\varepsilon(r)^2,$$

and it follows from (3.19) that

$$\log(\varepsilon(\tau)^2) \sim \log G(\tau) - \log(-\log G(\tau)), \tag{3.20}$$

where here and henceforth all asymptotic equivalences are understood to hold for $r \gg 1$. Hence

$$\varepsilon^2(\log G(\tau) - \log(-\log G(\tau))) \sim -G(\tau)$$

which in turn yields

$$G^{-1}\dot{G}\log G\left(1 - \frac{\log(-\log G)}{\log G} + \cdots\right) = 1,$$

and we obtain after integration

$$\frac{1}{2}(\log G)^2 = \tau\left(1 + \frac{\log(-\log G)}{\log G} + \cdots\right) + C$$

for some constant C. To the first term, the equality above gives $|\log G| \sim (2\tau)^{1/2}$, whence

$$G(\tau) \sim e^{-\sqrt{2\tau}} \quad \text{as } \tau \to \infty \tag{3.21}$$

up to some algebraic factor. From (3.21) and our choice of G, we deduce that

$$\int_\tau^\infty e^{\tau-s}\varepsilon(s)^2 ds \sim \varepsilon(\tau)^2. \tag{3.22}$$

it then follows from (3.22) and (3.17) that

$$\varepsilon(\tau)^2\log(\varepsilon(\tau)^2) = -G(\tau) + 4B\varepsilon(r)^2 + \cdots,$$

whence

$$\frac{\dot{G}}{G}(\log G - \log(-\log G) + 4B + \cdots) = 1.$$

Taking into account (3.21), we then see that

$$\begin{aligned}\frac{1}{2}\frac{d}{d\tau}((\log G)^2) &= \left(1 + \frac{1}{2\sqrt{\tau}}\left(\log\sqrt{2} + \frac{1}{2}\log\tau\right) + \frac{4B}{\sqrt{2\tau}} + \cdots\right)^{-1} \\ &= 1 - \frac{1}{2\sqrt{2}}\frac{\log\tau}{\sqrt{\tau}} - \frac{(\log\sqrt{2} + 4B)}{\sqrt{2\tau}} + \cdots.\end{aligned}$$

Setting $\alpha = \sqrt{2}(\log\sqrt{2} + 4B)$, we obtain

$$\frac{1}{2}(\log G)^2 = \tau - \frac{\tau^{\frac{1}{2}}}{\sqrt{2}}\log\tau - \alpha\tau^{\frac{1}{2}} + \cdots$$

which at once yields

$$G(\tau) \sim K_0 e^{-\sqrt{2\tau}}\tau^{\frac{1}{2\sqrt{\tau}}} \quad \text{with } K_0 = \sqrt{2}e^{4B}. \tag{3.23}$$

Plugging (3.23) and (3.20) into (3.19) and (3.18) follows.

We next set out to describe the way in which (1.3)–(1.5) are obtained. For the ease of presentation, we shall merely sketch the points that give rise to the different behaviors involved. To begin with, we continue to suppose that $N = 2$ and that Assumption 3.1 is in force. We now recast (3.5) in the form

$$\psi(y,\tau) = \sum_{k=0}^{\infty} a_k(\tau)\varphi_k(y) = \sum_{k=0}^{l} a_k(\tau)\varphi_k(y) + Q(y,\tau). \tag{3.24}$$

It is readily seen that formulas (3.6) read in this case as

$$\dot{a}_k = (1-k)a_k + \gamma\varepsilon(\tau)^2\langle\varphi_k, \delta(y)\rangle \quad \text{for } \tau \gg 1, \quad k = 0, 1, \ldots, l, \tag{3.25a}$$

$$Q_\tau = AQ + \gamma\varepsilon(\tau)^2\left(\delta(y) - \sum_{k=0}^{l}\langle\varphi_k, \delta(y)\rangle\varphi_k\right), \tag{3.25b}$$

$$\langle Q, \varphi_k\rangle = 0 \quad \text{for } k = 0, 1, \ldots, l. \tag{3.25c}$$

We now replace Assumption 3.2 by the following.

Assumption 3.3. The leading term in (3.24) as $\tau \gg 1$ is $a_l(\tau)\varphi_l(y)$; i.e., evolution in time of $\psi(y,\tau)$ is driven by the eigenfunction corresponding to the lth eigenvalue. Moreover, one then expects

$$\frac{d}{d\tau}(\varepsilon^2(\tau)) \sim (1-l)\varepsilon^2(\tau) \quad \text{as } \tau \to \infty, \tag{3.26a}$$

$$Q(y,\tau) \sim \gamma\varepsilon(\tau)^2 F(y) \quad \text{as } \tau \to \infty, \tag{3.26b}$$

where $F(y)$ satisfies

$$A_l F + \left(\delta(y) - \sum_{k=0}^{l}\langle\varphi_k, \delta(y)\rangle\varphi_k\right) = 0, k = 0, 1, \ldots, l, \tag{3.27a}$$

$$\langle F, \varphi_k\rangle = 0 \quad \text{for } k = 0, 1, \ldots, l \tag{3.27b}$$

and the operator A_l is given by

$$A_l F \equiv F_{yy} + \left(\frac{1}{y} - \frac{y}{2}\right)F_y + lF. \tag{3.27c}$$

Integrating (3.27), we obtain

$$F(y) = -\frac{1}{2\pi}\log y + \cdots \quad \text{for } \varepsilon(\tau) \ll y \ll 1. \tag{3.28}$$

We point out that no further details on the expansion (3.28) are required to derive the sought-for result (1.3). Arguing as before, we now obtain the following outer expansion for $\Phi(y,\tau)$:

$$\Phi(y,\tau) \sim \frac{y^2}{4} - \gamma\varphi_l(0)^2\int_\tau^\infty \varepsilon(s)^2 e^{(1-l)(\tau-s)}ds + \varepsilon(\tau)^2\left(-\frac{1}{2\pi}\log y + O(1)\right) \tag{3.29}$$

whenever $\varepsilon(\tau) \ll y \ll 1$ and $\tau \gg 1$. The inner expansion for $\Phi(y,\tau)$ is exactly that already obtained in (3.16). From (3.29) and (3.16), we deduce the matching condition

$$-\gamma\varphi_l(0)^2 \int_\tau^\infty \varepsilon(s)^2 e^{(1-l)(\tau-s)}ds = \frac{1}{2}\varepsilon^2(\tau)\log\varepsilon(\tau), \tag{3.30}$$

which yields now the following asymptotic behavior for $\varepsilon(\tau)$:

$$\varepsilon(\tau) \sim Ce^{(\frac{1-l}{2})\tau}\tau^{-\frac{1}{l-1}} \quad \text{as } \tau \to \infty \tag{3.31}$$

for some positive constant C.

Estimate (3.31) can be obtained from (3.30) by means of an argument similar to that leading from (3.17) to (3.18). Indeed, setting $\tau(s) = \varepsilon^2(s)e^{-(1-l)s}$ and observing that $4\gamma\varphi_l(0)^2 = 1$ (cf. (2.8)), we may rewrite (3.20) in the form

$$-\int_\tau^\infty r(s)ds = r(\tau)\log(\varepsilon^2(\tau)) = r(\tau)\left((1-l)\tau + \log(r(\tau))\right).$$

Hence

$$\frac{1}{l-1}\int_\tau^\infty r(s)ds = r(\tau)\left(r + O(\log(r(\tau)))\right) \quad \text{for } \tau \gg 1,$$

which can be integrated to yield $r(\tau) = Cr^{-l/(l-1)}$, whence

$$\varepsilon^2(\tau) = Ce^{(1-l)\tau}\tau^{-\frac{l}{l-1}},$$

and (1.3) follows.

We conclude this section by sketching the formal derivation of (1.4) and (1.5) in Theorem 1.2. Consider first the case of (1.4). From Assumptions 3.1 and 3.2 (with $\varepsilon^2(\tau)$ replaced by $\varepsilon^N(\tau)$ where appropriate), we obtain the following outer expansion for $\Phi(y,\tau)$:

$$\Phi(y,\tau) \sim \frac{y^2}{2N} - \gamma\sum_{k=0}^{l}\varphi_k^2(0)\int_\tau^\infty \varepsilon^N(s)e^{(1-k)(\tau-s)}ds + O\left(\frac{\varepsilon^N(\tau)}{y^{N-2}}\right) \tag{3.32}$$

whenever $\varepsilon(\tau) \ll y \ll 1$ and $\tau \gg 1$. The corresponding inner expansion is also obtained in the form

$$\Phi(y,\tau) = \varepsilon^2(\tau)w(\xi,\tau) \quad \text{with } \xi = \frac{y}{\varepsilon(\tau)},$$

where $w(\xi,\tau) \sim \bar{w}(\xi)$ for large τ and $\bar{w}$ solves

$$\bar{w}_{\xi\xi} + \left(\frac{N-1}{\xi}\right)\bar{w}_\xi = 1 \quad \text{for } \xi > 1,$$

$$\bar{w}(1) = \bar{w}_\xi(1) = 0$$

(compare with (3.14)). This now yields

$$\bar{w}(\xi) = \frac{\xi^2}{2N} - \frac{1}{2(N-2)} + \frac{\xi^{2-N}}{N(N-2)},$$

whereupon the following inner expansion for Φ follows:

$$\Phi(y,\tau) \sim \frac{y^2}{2N} - \frac{\varepsilon^2(\tau)}{2(N-2)} + \frac{1}{N(N-2)}\left(\frac{\varepsilon(\tau)}{y}\right)^{N-2} \quad \text{for } y \sim \varepsilon(\tau) \text{ and } \tau \gg 1. \tag{3.33}$$

Matching (3.32) and (3.33) gives the following equation for $\varepsilon(\tau)$:

$$\gamma \sum_{k=0}^{1} \varphi_k(0)^2 \int_{\tau}^{\infty} \varepsilon(s)^N e^{(1-k)(\tau-s)} ds = \frac{\varepsilon(\tau)^2}{2(N-2)}, \tag{3.34}$$

which yields $\varepsilon(\tau) \sim C\tau^{-1/(N-2)}$ and hence (1.4). Finally, (1.5) is obtained by guessing an outer expansion of the form

$$\Phi(y,\tau) \sim \frac{y^2}{2N} - \alpha e^{(1-l)\tau}\varphi_l(y). \tag{3.35}$$

This follows by neglecting the term $\chi_{\varepsilon(\tau)}$ in (3.2) and assuming that the lth mode dominates in the Fourier expansion for $\psi(y,\tau)$. Matching (3.35) with (3.33), (1.5) follows.

4. The topological argument. In this section we shall describe the basic approach towards a rigorous derivation of Theorems 1.1–1.3. For simplicity, we shall concentrate on the case where $N = 2$ and (1.2) holds and remark briefly on the remaining situations afterwards.

4.1. Obtaining (1.2) in Theorem 1.1. Let us define $\bar{\varepsilon}(\tau)$ as follows:

$$\bar{\varepsilon}(\tau) = Ke^{-\frac{\sqrt{2\tau}}{2}}\tau^{\frac{1}{4\sqrt{\tau}}-\frac{1}{4}}, \quad K \text{ given in (3.18).} \tag{4.1}$$

In another words, $\bar{\varepsilon}(\tau)$ is the leading part in the expected asymptotic behavior of the rescaled free boundary in this case. Fix now τ_0, τ_1 with $\tau_1 \geq \tau_0 \gg 1$ and consider functions $\varepsilon(\tau)$ such that the following estimates hold for some choice of $M > 1$.

$$\sup\left\{|\varepsilon(\tau) - \bar{\varepsilon}(s)|, \text{ where } \tau, s \in [\tau_0,\tau_1] \text{ and } |\tau - s| < \frac{1}{\tau}\right\} < M\varepsilon(\tau)\tau^{-\frac{3}{2}}, \tag{4.2a}$$

$$\frac{\bar{\varepsilon}(\tau)}{M} < \varepsilon(\tau) < M\bar{\varepsilon}(\tau) \quad \text{for } \tau \in [\tau_0,\tau_1]. \tag{4.2b}$$

We next recall that if $\Phi(y,\tau)$ is a (rescaled) solution of our problem (cf. (2.3)), then $\psi(y,\tau)$ given in (3.1) solves

$$\psi_\tau = A\psi + \chi_{\varepsilon(\tau)} \quad \text{for } y \in \mathbb{R}, \quad \tau > \tau_0, \tag{4.3a}$$

$$\psi(y,\tau_0) = \psi_0(y) \quad \text{at } \tau = \tau_0, \tag{4.3b}$$

where A is the linear operator in (2.5). We want to pick $\psi_0(y)$ above in a particular manner. Namely, we take

$$\psi_0(y) = \alpha_0\tilde{\varphi}_0(y) + \alpha_1\tilde{\varphi}_1(y) + \gamma\bar{\varepsilon}(\tau_0)^2 F(y), \tag{4.4}$$

where $F(y)$ is as in (3.9), and α_0, α_1, $\tilde{\varphi}_0$, and $\tilde{\varphi}_1$ will be selected presently. As a matter of fact, for $j = 0, 1$ functions $\tilde{\varphi}_j(y)$ will coincide with the eigenfunctions $\varphi_j(y)$ given in (2.6) for, say, $y \geq (\bar{\varepsilon}(\tau_0))^{1/2}$. The main point in selecting $\psi_0(y)$ in (4.4) is that we want it to match with the inner expansion (3.16) (with $\varepsilon(\tau)$ replaced by $\bar{\varepsilon}(\tau_0)$ there) at distances $y \sim \bar{\varepsilon}(\tau_0)^{1/2}$. This amounts essentially to imposing

$$\begin{aligned} &\alpha_0\varphi_0(0) + \alpha_1\varphi_1(y) + \gamma\bar{\varepsilon}(\tau_0)^2\left(-\frac{1}{2\pi}\log y + B + \cdots\right) \\ &\quad = \frac{\bar{\varepsilon}(\tau_0)^2}{2}\log\bar{\varepsilon}(\tau_0) - \frac{\bar{\varepsilon}(\tau_0)^2}{2}\log y - \frac{\bar{\varepsilon}(\tau_0)^2}{4}, \end{aligned}$$

whence

$$\alpha_0\varphi_0 + \alpha_1\varphi_0(0) + \left(B + \frac{1}{4}\right)\gamma\bar{\varepsilon}(\tau_0)^2 = \frac{\bar{\varepsilon}(\tau_0)^2}{2}\log\bar{\varepsilon}(\tau_0) \tag{4.5}$$

so that

$$|\alpha_0| + |\alpha_1| = O(\bar{\varepsilon}(\tau_0)^2|\log\bar{\varepsilon}(\tau_0)|). \tag{4.6}$$

We have yet to determine what kind of modification is to be performed on the φ_k's near the origin for $k = 0, 1$. To ascertain this point, we observe that if no change were done at all, we would have that

$$\psi_0(y) \sim -\frac{\gamma\bar{\varepsilon}(\tau_0)^2}{2\pi}\log y \quad \text{as } y \to 0.$$

To remove such singularity, we just redefine the φ_k's near $y = 0$ as follows:

$$\alpha_0\tilde{\varphi}_0 + \alpha_1\tilde{\varphi}_1(y) = \frac{\gamma\bar{\varepsilon}(\tau_0)^2}{2\pi} + o(\bar{\varepsilon}(\tau)^2\log y) \quad \text{as } y \to 0. \tag{4.7}$$

Notice that relations (4.4)–(4.7) are compatible and allow for many possible choices of α_k, $\tilde{\varphi}_k$ for $k = 0.1$. Bearing in mind our previous arguments, we now introduce the following notation:

(4.8)

> Let τ_0, τ_1 be such that $\tau_1 \geq \tau_0 \gg 1$, and let μ be a given number such that $0 < \mu \leq 1$. We shall say that a solution $\psi(y,\tau)$ of (4.3a) which is defined for $\tau_0 \leq \tau \leq \tau_1$ belongs to the class $A(\tau_0, \tau_1, \mu)$ if there exists a constant M such that $|\psi(y,\tau)| < M(1 + y^2)$ for $y \in \mathbb{R}$ and $\tau \in [\tau_0, \tau_1]$ and conditions (4.2) are satisfied with M replaced by $M\mu$ there.

We shall say that $\psi(y,\tau) \in \overline{\mathcal{A}(\tau_0, \tau_1, \mu)}$ if it satisfies those conditions describing membership in the class $\mathcal{A}(\tau_0, \tau_s, \mu)$ when strict inequalities are replaced by the symbol $\leq$.

For $k = 0, 1$, let us now define

$$l_k(\alpha_0, \alpha_1; \tau) = \langle\psi(y, \tau; \alpha_0, \alpha_1), \varphi_k\rangle + \int_\tau^\infty e^{(1-k)(\tau - s)}\langle\chi_{\bar{\varepsilon}}, \varphi_k\rangle ds, \tag{4.9}$$

where $\psi(y, \tau; \alpha_0, \alpha_1)$ is the solution of (4.3a) such that $\psi(y, \tau_0; \alpha_0, \alpha_1) = \psi_0(y)$, $\psi_0(y)$ being a function satisfying (4.4)–(4.7) above.

The following result is a crucial ingredient in the proof of (1.2) in Theorem 1.1.

PROPOSITION 4.1. *Assume that $M > 0$ is large enough and let $\psi(y,\tau)$ be the solution of* (4.3)*, where $\psi(y,\tau_0) \equiv \psi_0(y)$ is such that* (4.4)–(4.7) *hold. Suppose also that*

$$\psi(y,\tau) \in \overline{\mathcal{A}(\tau_0,\tau_1,1)} \tag{4.10}$$

for some τ_1, τ_0 so that $\tau_1 > \tau_0 \gg 1$. Then if

$$l_k(\alpha_0,\alpha_1,\tau_1) = 0 \quad \textit{for } k = 0,1 \tag{4.11}$$

(*cf.* (4.9) *above*)*, one has that*

$$\psi(y,\tau) \in \mathcal{A}\left(\tau_0,\tau_1,\frac{1}{2}\right).$$

We shall prove Proposition 4.1 in sections 5 and 6, which contain most of the technical aspects of this paper. To keep the flow of the main arguments here, we will assume that the proposition holds true and continue with the derivation of (1.2). Let $\alpha = (\alpha_0,\alpha_1)$ be any pair of real numbers and set $l(\alpha_0,\alpha_1;\tau) = (l_0(\alpha_0,\alpha_1;\tau), l_1(\alpha_0,\alpha_1;\tau))$, where for $k = 0,1, l_k$ is defined as the right-hand side of (4.9). Let τ_1, τ_0 be such that $\tau_1 \geq \tau_0$ and define $\mathcal{U}(\tau_0,\tau_1) \subset \mathbb{R}^2$ as the open set consisting of all points $(\alpha_0,\alpha_1) \in \mathbb{R}^2$ such that the corresponding solution $\psi(y,\tau)$ of (4.3)–(4.8) satisfies that $\psi(y,\tau) \in \mathcal{A}(\tau_0,\tau_1,1)$. From our previous arguments, it follows that we may select an initial value $\psi(y,\tau_0) = \psi_0(y)$ in (4.3b) so that

$$\psi(y,\tau_0) \in \mathcal{A}\left(\tau_0,\tau_0,\frac{1}{2}\right)$$

and there exists a unique solution of $l(\alpha_0,\alpha_1,\tau_0) = 0$. Indeed, by (4.9) one has that

$$l_k(\alpha_0,\alpha_1;\tau_0) = \alpha_k + \delta(\alpha,\tau_0)(|\alpha_0| + |\alpha_1|) + O(\bar{\varepsilon}(\tau_0)^2) \quad \text{for } k = 0,1,$$

where $\delta(\alpha,\tau_0) \to 0$ as $\tau_0 \to \infty$, uniformly for $|\alpha| = |\alpha_0| + |\alpha_1|$ bounded, and the last term on the right above may be assumed to be independent on α. On the other hand, we may always suppose $l = (l_0,l_1)$ to be differentiable with respect to α_0,α_1 by means of a suitable choice of the initial value $\psi_0(y)$. We shall assume henceforth that $\varphi_0(y)$ satisfies such a condition. It then turns out that for $k = 0,1$ equation $l_k(\alpha_0,\alpha_1;\tau_0) = 0$ has a unique solution α_k such that

$$\alpha_k = O(\bar{\varepsilon}(\tau_0)^2).$$

As a matter of fact, one then has that

$$d(l,\mathcal{U}(\tau_0,\tau_0);0) = 1,$$

where for $\tau \geq \tau_0$, $d(l,\mathcal{U}(\tau_0,\tau);0)$ denotes the topological degree of the mapping l in the set $\mathcal{U}(\tau_0,\tau)$ at the value zero.

Now assume that $\mathcal{U}(\tau_0,\tau) \neq \phi$ for any $\tau \in [\tau_0,\tau_1]$ with $\tau_1 > 0$ and denote by $\partial\mathcal{U}(\tau_0,\tau)$ the boundary of the open set $\mathcal{U}(\tau_0,\tau)$. We notice that if $l \neq 0$ on $U(\partial\mathcal{U}(\tau_0,\tau))$ for $\tau_0 \leq \tau \leq \tau_1$, the $d(l,\mathcal{U}(\tau_0,\tau);0) = d(l,\mathcal{U}(\tau_0,\tau_0);0)$ for any such τ. It then follows from standard continuous dependence results that

$$\mathcal{U}(\tau_0,\tau_1) \neq \phi$$

and

$$d(l,\mathcal{U}(\tau_0,\tau_1);0)=1$$

for any $\tau_1>\tau_0$ such that $(\tau_1-\tau_0)$ is sufficiently small. We next claim that

$$d(l,\mathcal{U}(\tau_0,\tau);0)=1 \tag{4.12a}$$

for any $\tau>\tau_0$ as far as

$$\mathcal{U}(\tau_0,\tau)\neq\phi. \tag{4.12b}$$

Indeed, suppose that there exists a first time $\tau>\tau_0$ when (4.12a) fails but (4.12b) holds true. In view of our previous remark, there must be a point $\beta=(\beta_0,\beta_1)\in\partial\mathcal{U}(\tau_0,\tau)$, where $l(\beta)=0$, and clearly $\psi(y,\tau;\beta_0,\beta_1)\in\overline{\mathcal{A}(\tau_0,\tau;1)}$. We then use Proposition 4.1 to deduce that $\beta\in\mathcal{U}(\tau_0,\tau)$, which is a contradiction.

We further observe that

$$\mathcal{U}(\tau_0,\tau)\neq\phi\quad\text{for any }\tau>\tau_0, \tag{4.13}$$

$$\text{provided that }\tau_0\gg 1.$$

To check (4.13), we define $\tau^*=\sup\{\tau:\mathcal{U}(\tau_0,\tau)\neq\phi\}$. We already know that $\tau^*>\tau_0$. Assume now that $\tau^*<\infty$. By (4.12), we may select a sequence of times $\{\tau_n\}$ increasing to τ^* and a sequence $\{\alpha_n\}=\{(\alpha_{0n},\alpha_{1n})\}$ such that $l(\alpha_{0n},\alpha_{1n};\tau_n)=0$ and $\alpha_n\in\mathcal{U}(\tau_0,\tau_n)$. Since $\mathcal{U}(\tau_0,\tau_{n+1})\subset\mathcal{U}(\tau_0,\tau_n)$, one has that $\{\alpha_n\}$ is bounded. Therefore, a subsequence (still denoted by $\{\alpha_n\}$) exists which converges to some point $\alpha^*=(\alpha_0^*,\alpha_1^*)$. It then turns out that $l(\alpha_0^*,\alpha_1^*;\tau^*)=0$ and hence by Proposition 4.1 the corresponding function $\psi(y,\tau;\alpha_0^*,\alpha_1^*)$ remains at the interior of $\mathcal{A}(\tau_0,\tau^*;1)$; this is the point where restriction $\tau_0\gg 1$ needs to be imposed on (4.13). By continuous dependence results, ψ would also remain at the interior of $\mathcal{A}(\tau_0,\tau^*+\delta;1)$ for some $\delta>0$, thus contradicting the definition of τ^*.

We are now prepared to detail the argument leading to the existence of the solutions referred to in Theorems 1.1 and 1.3. Take a sequence $\{\tau_n\}$ such that $\tau_1>\tau_0$ and $\lim_{n\to\infty}\tau_n=\infty$. For any such n, $\mathcal{U}(\tau_0,\tau_n)\neq\phi$, and we may select $\alpha_n=(\alpha_{0n},\alpha_{1n})$ such that $l(\alpha_{0n},\alpha_{1n};\tau_n\}=0$.

Let $\psi_n(y,\tau)\equiv\psi_n(y,\tau;\alpha_{0n},\alpha_{1n})$ be the solution of (4.3a) with initial value $\psi_n(y,\tau_0)=\psi_0(y;\alpha_{0n},\alpha_{1n})$ satisfying (4.4)–(4.8). By Proposition 4.1, we have that $\psi_n(y,\tau)\in\mathcal{A}(\tau_0,\tau_n;\frac{1}{2})$. Since the sequence $\{\alpha_n\}$ is bounded, there exists a subsequence (still denoted by $\{\alpha_n\}$) and a value $\bar\alpha=(\bar\alpha_0,\bar\alpha_1)$ such that $\lim_{n\to\infty}\alpha_n=\bar\alpha\in\mathcal{U}(\tau_0,\tau_0)$. It then turns out that function $\psi(y,\tau;\bar\alpha_0,\bar\alpha_1)$, solution of (4.3a) with initial value $\psi(y,\tau_0;\bar\alpha_0,\bar\alpha_1)$, provides a sought-for solution satisfying (1.2), and the proof is concluded under our current assumptions.

4.2. The remaining cases. To derive (1.3) in Theorem 1.1, we just repeat our previous argument with the following modifications. First we replace $\bar\varepsilon(\tau)$ in (4.1) by

$$\bar\varepsilon(\tau)=Ce^{(1-l)\tau}\tau^{-\frac{1}{l-1}},$$

where, as in the statement of the theorem, l is any number larger than or equal to two and C is any positive constant. Instead of making use of (4.4), we now define $\psi_0(y)$ by

$$\psi_0(y)=\sum_{k=0}^{l}\alpha_k\tilde\varphi_k(y)-\frac{C}{2\pi}\log y$$

and replace condition (4.11) in Proposition 4.1 by

$$l_k(\alpha_0, \alpha_1, \ldots, \alpha_0; \tau_1) = 0 \quad \text{for } k = 0, 1, \ldots, l,$$

where the l_k's are defined as in (4.9), except that here we allow l_k to depend on all parameters $\alpha_0, \alpha_1, \ldots, \alpha_l$.

The cases corresponding to dimensions $N \geq 3$ are similar. For instance, to obtain (1.4) (resp. (1.5)), we define $\bar{\varepsilon}(\tau)$ as follows:

$$\bar{\varepsilon}(\tau) = B\tau^{-\frac{1}{N-2}} \quad (\text{resp. } \bar{\varepsilon}(\tau) = Ce^{(1-\frac{l}{2})\tau}),$$

where $B > 0$ is a fixed constant which depends on N and can be determined from (3.34) and C is any given constant. We now have to redefine the φ_k's near $y = 0$ in order to remove singularities of the type $y^{-(N-2)}$ instead of logarithmic ones. A straightforward modification of the previous approach yields then the desired results.

5. Derivation of (1.2): Analysis of the outer region. We now set out to provide the details required to justify the picture given in sections 3 and 4. To this end, we shall concentrate on proving Proposition 4.1 to highlight those modifications required to obtain (1.3)–(1.5). From now on we shall thus assume that $N = 2$ and start by considering solutions to the equation satisfied by $\psi(y, \tau)$ given in (3.2); i.e.,

$$\psi_\tau = A\psi + \chi_{\varepsilon(\tau)} \quad \text{for } \tau > \tau_0, y \in \mathbb{R} \tag{5.1}$$

with initial condition (4.3b), where $\psi_0(y)$ satisfies (4.4)–(4.8) and operator A is given in (2.5). We shall compare solutions to (5.1) and (4.3b) with those to the auxiliary equation

$$W_\tau = AW + \gamma\varepsilon(\tau)^2\delta(y) \quad \text{for } \tau > \tau_0, y \in \mathbb{R} \tag{5.2}$$

with the same initial condition at $\tau = \tau_0$. Solutions of (5.1) can be represented in the form

$$\psi(y, \tau) = a_0(\tau)\varphi_0(y) + a_1(\tau)\varphi_1(y) + E(y, \tau), \tag{5.3}$$

where $E(y, \tau)$ satisfies

$$E_\tau = \Delta E - \frac{1}{2}y\nabla E + E + (\chi_{\varepsilon(\tau)} - \langle\varphi_0, \chi_{\varepsilon(\tau)}\rangle\varphi_0 - \langle\varphi_1\chi_{\varepsilon(\tau)}\rangle\varphi_1) \tag{5.4}$$

and $\langle E, \varphi_k\rangle = 0$ for $k = 0, 1$. Bearing in mind (5.2), we shall also consider solutions $Q(y, \tau)$ to the equation

$$\begin{aligned} Q_\tau = \Delta Q - \frac{1}{2}y\nabla Q + Q \\ + \gamma\varepsilon(\tau)^2(\delta(y) - \langle\varphi_0, \delta(y)\rangle\varphi_0 - \langle\varphi_1, \delta(y)\rangle\varphi_1) \end{aligned} \tag{5.5}$$

such that $\langle Q, \varphi_k\rangle = 0$ for $k = 0, 1$. We then have the following lemma.

LEMMA 5.1. *Let $\tau_0 > 0$ be fixed. Then the solution $Q(y, \tau_0)$ of (5.5) which is defined for $\tau > \tau_0$ and satisfies $Q_0(y, \tau_0) = 0$ is given by*

$$Q_0(y, \tau) = \gamma \int_{\tau_0}^{\tau} K(y, \tau - s)e^{\tau - s}\varepsilon(s)^2 ds, \tag{5.6a}$$

where

$$\begin{aligned}K(y,\tau) &= (4\pi(1-e^{-\tau}))^{-1}\left(\exp\left(-\frac{y^2e^{-\tau}}{4(1-e^{-\tau})}\right)\right.\\ &\left.-\sum_{j=0}^{1}\varphi_j\left\langle\varphi_j,\exp\left(-\frac{y^2e^{-\tau}}{4(1-e^{-\tau})}\right)\right\rangle\right).\end{aligned}\tag{5.6b}$$

Proof. For convenience, we shall dispense with the subscript in $Q_0(y,\tau)$. Differentiating three times with respect to the y variables in (5.5) yields

$$\begin{aligned}\frac{\partial}{\partial\tau}Q_{i,j,k} &= \Delta Q_{i,j,k}-\frac{1}{2}y\nabla Q_{i,j,k}-\frac{1}{2}Q_{i,j,k}+\gamma\varepsilon(\tau)^2(\delta(y))_{i,j,k}\\ &= A_*Q_{i,j,k}-\frac{1}{2}Q_{i,j,k}+\gamma\varepsilon(\tau)^2(\delta(y))_{i,j,k}\end{aligned}$$

in an appropriate weak sense. Using a variation of constants formula in the equation above and denoting by S_* the semigroup generated by A_*, we obtain that

$$\begin{aligned}Q_{i,j,k}(y,\tau) &= \gamma\int_{\tau_0}^{\tau}e^{-(\frac{\tau-s}{2})}S_*(\tau-s)\left(\varepsilon(s)^2\frac{\partial^3(\delta(\xi))}{\partial\xi_i\partial\xi_j\partial\xi_k}\right)ds = -\gamma\int_{\tau_0}^{\tau}e^{-\frac{(\tau-s)}{2}}\varepsilon(s)^2\\ &\cdot\int_{\mathbb{R}^2}\frac{\partial^3}{\partial\xi_i\partial\xi_j\partial\xi_k}\left((4\pi(1-e^{-(\tau-s)}))^{-1}\exp\left(-\frac{(ye^{-(\frac{\tau-s}{2})}-\xi)^2}{4(1-e^{-(\tau-s)})}\delta(\xi)\right)d\xi\right)ds\\ &= \gamma\frac{\partial^3}{\partial y_i\partial y_j\partial j_k}\int_{\tau_0}^{\tau}e^{\tau-s}\varepsilon(s)^2\left((4\pi(1-e^{-(\tau-s)}))^{-1}\exp\left(-\frac{(ye^{-(\frac{\tau-s}{2})})^2}{4(1-e^{-(\tau-s)})}\right)\right)ds.\end{aligned}$$

Integrating now three times with respect to the y variables and imposing $\langle Q_1\varphi_0\rangle = \langle Q_1\varphi_1\rangle = 0$, the result follows. □

We shall elaborate a bit on the formulas in (5.6). To begin with, we observe that

$$\left\langle\varphi_0,\exp\left(-\frac{y^2e^{-(\tau-s)}}{4(1-e^{-(\tau-s)})}\right)\right\rangle = C(1-e^{-(\tau-s)})\quad\text{for some } C>0,\tag{5.7a}$$

$$\left\langle\varphi_1,\exp\left(-\frac{y^2e^{-(\tau-s)}}{4(1-e^{-(\tau-s)})}\right)\right\rangle = (a_0+a_1(1-e^{-(\tau-s)}))(1-e^{-(\tau-s)})\tag{5.7b}$$

for some constants a_0 and a_1. To check (5.7a), we simply notice that

$$\begin{aligned}\left\langle\varphi_0,\exp\left(-\frac{y^2e^{-(\tau-s)}}{4(1-e^{-(\tau-s)})}\right)\right\rangle &= c_0\int_{\mathbb{R}^2}\exp\left(-\frac{y^2}{4}-\frac{y^2e^{-(\tau-s)}}{4(1-e^{-(\tau-s)})}\right)dy\\ &= c_0(1-e^{-(\tau-s)})\int_{\mathbb{R}^2}e^{-r^2}dr,\end{aligned}$$

where c_0 is given in (2.6b). The proof of (5.7b) is similar and will therefore be omitted.

Assume now that $\tau>\tau_0+1$. We may then split the integral term in (5.6a) in the form

$$Q_0(y,\tau) = \int_{\tau_0}^{\tau-1}(\;)+\int_{\tau-1}^{\tau}(\;)\equiv Q_{0,1}(y,\tau)+Q_{0,2}(y,\tau).\tag{5.8}$$

To estimate $Q_{0,2}$, we proceed to examine the quantity

$$J \equiv \gamma \int_{\tau-1}^{\tau} (4\pi(1-e^{-(\tau-s)}))^{-1} \exp\left((\tau-s) - \frac{y^2 e^{-(\tau-s)}}{4(1-e^{-(\tau-s)})}\right) \varepsilon^2(s)ds. \tag{5.9}$$

Setting $\eta = y^2 e^{-(\tau-s)}(4(1-e^{-(\tau-s)}))^{-1}$, it follows that $e^{\tau-s} = 1+\frac{y^2}{4\eta}$ and $d\eta = 4^{-1}y^2 e^{-(\tau-s)}(1-e^{-(\tau-s)})^{-2}ds$. If we now write $f(y) = e\theta(y)$ with $\theta(y) = y^2e^{-1}(4(1-e^{-1}))^{-1}$, we easily see that

$$J = \frac{\gamma}{\pi}\int_{\theta(y)}^{\infty} \varepsilon(s)^2 \left(1+\frac{y^2}{4\eta}\right)^2 y^{-2}e^{-\eta}d\eta = \frac{\gamma}{4\pi}\int_{f(y)}^{\infty} \varepsilon(s)^2\left(1+\frac{y^2}{4\eta}\right)e^{-\eta}\eta^{-1}d\eta,$$

where $s = \tau - \log(1+\frac{y^2}{4\eta})$.

Now we shall pay attention to the term $S(\tau-\tau_0)Q(y,\tau_0)$, where $Q(y,\tau_0)$ is given by

$$\begin{aligned} Q(y,\tau_0) &= \gamma\bar{\varepsilon}(\tau_0)^2F(y) + \sum_{j=0}^{1}\alpha_j\left(\tilde{\varphi}_j - \sum_{k=0}^{1}\varphi_k\langle\varphi_k,\tilde{\varphi}_j\rangle\right) \\ &\equiv \gamma\bar{\varepsilon}(\tau_0)^2F(y) + R(y), \end{aligned} \tag{5.10}$$

where α_j, φ_j, and $\tilde{\varphi}_j$ are as in (4.4). Notice that $\langle Q(\cdot,\tau_0),\varphi_k\rangle = 0$ for $k = 0, 1$. Moreover, one has that

$$R(y) = \sum_{j=0}^{1}\alpha_j(\tilde{\varphi}_j - \varphi_j) + \sum_{j=0}^{1}\alpha_j\left(\sum_{k=0}^{1}\varphi_k(\delta_{j,k} - \langle\varphi_k,\tilde{\varphi}_j\rangle)\right),$$

where, as customary, $\delta_{j,k} = 1$ if $j = k$ and $\delta_{j,k} = 0$ otherwise. In view of (4.4)–(4.7), it holds that

$$|R(y)| \le C\left(\bar{\varepsilon}(\tau_0)^2\left|\log\left(\frac{y}{\bar{\varepsilon}(\tau_0)}\right)\right|\chi_{[y\le\bar{\varepsilon}(\tau_0)^{1/2}]} + \bar{\varepsilon}(\tau_0)^3\tau_0(1+y^2)\right).$$

Using the explicit kernel for the semigroup $S(\tau)$, we then derive

$$|S(\tau-\tau_0)R| \le C\bar{\varepsilon}(\tau_0)^3\tau_0^{\frac{3}{2}}(1+y^2) \quad \text{for } \tau_0 \le \tau \le \tau_0+1 \text{ and } y \ge \bar{\varepsilon}(\tau_0)^{\frac{1}{4}}. \tag{5.11a}$$

On the other hand, by regularizing properties of $S(\tau)$, we obtain

$$\begin{aligned} &|S(\tau-\tau_0)R| \le C\bar{\varepsilon}(\tau)^{3-\chi}(1+y^2) \quad \text{for } \tau \ge \tau_0+1, \\ &\qquad y \ge \bar{\varepsilon}(\tau_0)^{\frac{1}{4}}, \quad \text{and some } \chi \in (0,1). \end{aligned} \tag{5.11b}$$

Summing up, we have obtained the following lemma.

Lemma 5.2. *Let $\tau^* = \max\{\tau_0, \tau-1\}$ and let $Q(y,\tau)$ be the solution of (5.5) for $\tau > \tau_0$ such that $Q(y,\tau_0)$ is given by (5.10). We then have that*

$$\begin{aligned} Q(y,\tau) = {} & \gamma\bar{\varepsilon}(\tau_0)^2S(\tau-\tau_0)F(y) + \gamma\int_{\tau_0}^{\tau^*}K(y,\tau-s)e^{\tau-s}\varepsilon(s)^2ds \\ & + \gamma\int_{\tau^*}^{\tau}(A_1+A_2(1-e^{-(\tau-s)}))e^{\tau-s}\varepsilon(s)^2ds + \frac{\gamma}{\pi}\int_{\Sigma}^{\infty}\varepsilon(s)^2e^{-\eta}\eta^{-1}d\eta \\ & + \frac{\gamma}{\pi}\int_{\Sigma}^{\infty}\varepsilon(s)^2e^{-y}\eta^2(4\eta^2)^{-1}d\eta + O\left(\frac{\bar{\varepsilon}(\tau)^2}{\tau}\right)(1+y^2) \end{aligned} \tag{5.12}$$

in regions where $y \geq \bar{\varepsilon}(\tau_0)^{1/4}$ *and* $\tau > \tau_0$. *Here* A_1, A_2 *are some positive constants,* $\sum = y^2(4(e^\beta - 1))^{-1}$, *where* $\beta = \max\{\tau_0, \tau - 1\}$ *and in the last two integrals above,* $s = \tau - \log(1 + \frac{y^2}{4\eta})$.

We now proceed to estimate the difference $(E - Q)$, where E and Q are solutions of (5.4) and (5.5), respectively. To this end, we set

$$Z = E - Q, \quad g = \chi_{\varepsilon(\tau)} - \gamma\varepsilon(\tau)^2\delta(y) \tag{5.13}$$

so that Z satisfies

$$\begin{aligned} Z_\tau &= \Delta Z - \frac{1}{2}\eta\nabla Z + Z + (g - \langle\varphi_0, g\rangle\varphi_0 - \langle\varphi_1, g\rangle\varphi_1) \\ &\equiv AZ + (g - \langle\varphi_0, g\rangle\varphi_0 - \langle\varphi_1, g\rangle\varphi_1) \equiv AZ + h(y, \tau) \end{aligned} \tag{5.14a}$$

We shall consider equation (5.14a) for values $\tau > \tau_0 \gg 1$. At $\tau = \tau_0$, we impose $E(y, \tau_0) = Q(y, \tau_0)$ so that

$$Z(y, \tau_0) = 0. \tag{5.14b}$$

We then have that the solution to (5.14) can be written in the form

$$Z(y, \tau) = \int_{\tau_0}^{\tau} S(\tau - s)h(\cdot, s)ds \equiv \int_{\tau_0}^{\tau} L(y, \tau - s; s)ds, \tag{5.15a}$$

where

$$\begin{aligned} L(y, \tau - s; s) &= (4\pi(1 - e^{-(\tau - s)}))^{-1} \\ &\cdot \int_{\mathbb{R}^2} \exp\left(-\frac{(ye^{-(\frac{\tau - s}{2})} - \xi)^2}{4(1 - e^{-(\tau - s)})}\right) h(\xi, s)d\xi. \end{aligned} \tag{5.15b}$$

Without loss of generality, we may assume $\tau > \tau_0 + 1$. We then split Z in the form

$$Z(y, \tau) = \int_{\tau_0}^{\tau - 1} L\,ds + \int_{\tau - 1}^{\tau} L\,ds. \tag{5.16}$$

Then the following lemma holds.

Lemma 5.3. *There exists* $C > 0$ *such that*

$$\int_{\tau_0}^{\tau - 1} |L|ds \leq C\int_{\tau_0}^{\tau - 1} e^{-(\tau - s)}\varepsilon(s)^4 ds \tag{5.17}$$

uniformly on bounded sets $|y| \leq R < \infty$.

Proof. To begin with, we observe that

$$\begin{aligned} |\langle\varphi_0, g(\cdot, s)\rangle| &= \left|\int_{\mathbb{R}^2} \varphi_0(\chi_{\varepsilon(s)} - \gamma\varepsilon(s)^2\delta(y))e^{-y^2/4}dy\right| \\ &= \left|\int_{|y| \leq \varepsilon(s)} \varphi_0(e^{-y^2/4} - 1)dy\right| \leq C\int_{|y| \leq \varphi(s)} r^2 dr \leq C\varepsilon(s)^4 \end{aligned}$$

and a similar bound is easily obtained for $|\langle\varphi_1, g\rangle|$. We thus have that

$$|\langle\varphi_0, g(\cdot,s)\rangle| + |\langle\varphi_1, g(\cdot,s)\rangle| \le C\varepsilon(s)^4 \tag{5.18}$$

for some $C > 0$.

Recalling that $\tau - s \ge 1$ under our current assumptions, we now consider the term

$$I = \int_{\tau_0}^{\tau-1} (4\pi(1-e^{-(\tau-1)}))^{-1} \int_{\mathbb{R}^2} \exp\left(-\frac{(ye^{-(\frac{\tau-s}{2})}-\xi)^2}{4(1-e^{-(\tau-s)})}\right) g(\xi,s)ds$$

$$\equiv \int_{\tau_0}^{\tau-1} S(\tau-s)g(\cdot,s)ds \equiv \int_{\tau_0}^{\tau-1} S\left(\frac{1}{2}\right) S\left(\tau-s-\frac{3}{4}\right) S\left(\frac{1}{4}\right) g(\cdot,s)ds. \tag{5.19}$$

We claim that

$$\left\| S\left(\frac{1}{4}\right) g(\cdot,s)\right\| \le C\varepsilon(s)^4 \quad \text{for some } C > 0. \tag{5.20}$$

Let us assume (5.20) for the moment and continue. One then may use classical regularizing effects to derive that

$$\left\| S\left(\tau-s-\frac{3}{4}\right)\left(S\left(\frac{1}{4}\right) g(\cdot,s)\right)\right\| \le Ce^{-(\tau-s)}\varepsilon(s)^4. \tag{5.21}$$

Finally, a standard Sobolev imbedding yields

$$\left| S\left(\frac{1}{2}\right)\left(S\left(\tau-s-\frac{3}{4}\right) S\left(\frac{1}{4}\right) g(\cdot,s)\right)\right| \le Ce^{-(\tau-s)}\varepsilon(s)^4. \tag{5.22}$$

Putting together (5.18)–(5.22), estimate (5.17) follows. The proof will thus be complete as soon as (5.20) has been obtained. To derive this last result, we make use of a duality argument. Let $\varphi(y)$ be any radial function in $L^2(\mathbb{R})$, and consider the integral

$$J = \int_{\mathbb{R}^2} g(\xi,s)\left(\int_{\mathbb{R}^2} e^{-(y-\xi)^2}\varphi(y)dy\right)d\xi \equiv \int_{\mathbb{R}^2} g(\xi,s)G(\xi)d\xi.$$

Recalling the arguments leading to (5.18), one readily sees that

$$\begin{aligned}\int_{\mathbb{R}^2} g(\xi,s)G(\xi)d\xi &= \int_{\mathbb{R}^2} (\chi_{\varepsilon(s)} - \gamma\varepsilon(s)^2\delta(y))G(\xi)d\xi \\ &= \int_{|y|\le\varepsilon(s)} (G(\xi)-G(0))d\xi \le C\varepsilon(s)^4,\end{aligned}$$

whereupon (5.20) follows. □

Our next result reads as follows.

Lemma 5.4. *There exists $C > 0$ such that*

$$\int_{\tau-1}^{\tau} |L(y,\tau-s;s)|ds \le C\left(\int_{\tau-1}^{\tau} \varepsilon(s)^4 ds + |y|\int_0^1 \varepsilon(\tau-s)^3 s^{-2} e^{-\frac{Cy^2}{s}} ds\right). \tag{5.23}$$

Proof. Set

$$w(y,\tau-s;\xi)=\exp\left(-\frac{(ye^{-(\frac{\tau-s}{2})}-\xi)^2}{4(1-e^{-(\tau-s)})}\right). \tag{5.24a}$$

We can then readily check that

$$\int_{\tau-1}^{\tau}L(y,\tau-s;s)ds=\int_{\tau-1}^{\tau}(4\pi(1-e^{-(\tau-s)}))^{-1}e^{\tau-s}M(y,\tau-s)ds, \tag{5.24b}$$

where

$$\begin{aligned} M(y,\tau-s)=\int_{\mathbb{R}^2}&(w(y,\tau-s;\xi)-\langle\varphi_0,w(y,\tau-s;\xi)\rangle\varphi_0\\ &-\langle\varphi_1,w(y,\tau-s;\xi)\rangle\varphi_1)g(\xi,s)d\xi. \end{aligned} \tag{5.24c}$$

$$\frac{y^2}{4}+\frac{(ye^{-(\frac{\tau-s}{2})}-\xi)^2}{4(1-e^{-(\tau-s)})}=\frac{1}{4(1-e^{-(\tau-s)})}\left((y-\xi e^{-(\frac{\tau-s}{2})})^2+\xi^2(1-e^{-(\tau-s)})\right).$$

A quick computation then reveals that

$$\langle\varphi_0,w(y,\tau-s;\xi)\rangle\le Ce^{-\frac{y^2}{4}}(1-e^{-(\tau-s)}),$$

and a similar result holds when we replace φ_0 by φ_1 above. Recalling the argument leading to (5.18), we then have that

$$\int_{\mathbb{R}^2}|(\langle\varphi_0,w\rangle\varphi_0+\langle\varphi_1,w\rangle\varphi_1)|\,|g(\cdot,s)|ds\le C(1-e^{-(\tau-s)})\varepsilon(s)^4. \tag{5.25}$$

Now consider the integral

$$J\equiv\int_{\tau-1}^{\tau}(4\pi(1-e^{-(\tau-s)}))^{-1}e^{\tau-s}\left(\int_{\mathbb{R}^2}w(y,\tau-s;\xi)g(\xi,s)d\xi\right)ds. \tag{5.26}$$

Since $\int_{\mathbb{R}^2}g(\xi,s)d\xi=0$, it holds that

$$J=\int_{\tau-1}(4\pi(1-e^{-(\tau-s)}))^{-1}e^{\tau-s}\left(\int_{\mathbb{R}^2}(w(y,\tau-s;\xi)-w(y,\tau-s;0))g(\xi,s)d\xi\right)ds,$$

whence

$$\begin{aligned} |J|\le C\textstyle\int_{\tau-1}^{\tau}&(4\pi(1-e^{-1(\tau-s)}))^{-1}\\ &\cdot\left(\int_{|\xi|\le\varepsilon(s)}|w(y,\tau-s;\xi)-w(y,\tau-s;0)|d\xi\right)ds. \end{aligned} \tag{5.27}$$

We now observe that for any real numbers a and b,

$$|e^{-a^2}-e^{-(a-b)^2}|\le Ce^{-a^2/2}|a|\,|b|\quad\text{for some } C>0. \tag{5.28}$$

To check (5.28), we consider first the case where $|a|\ge|b|$. Then if $|a|\,|b|\le 1$, we have that $|1-e^{2ab-b^2}|\le C|a|\,|b|$, whereas if $|a|\,|b|>1$, $|e^{-a^2}-e^{-(a-b)^2}|\le$

$Ce^{-a^2/2} \leq Ce^{-a^2/2}|a|\,|b|$. When $|a| < |b|$, we simply select $\mu > 0$ large enough and observe that if $|a|\,|b| < \mu$, then $|1 - e^{2ab-b^2}| \leq C|a|\,|b|$ for some $C = C(\mu) > 0$, whereas for $|a|\,|b| > \mu$ one has that $2ab - b^2 \geq -\frac{b^2}{2}$ and hence $|1 - e^{2ab-b^2}| \leq C|a|\,|b|$. Having shown that (5.28) holds, we now take advantage of that inequality (with $a = ye^{-\frac{(\tau-s)}{2}}(4(1 - e^{-(\tau-s)}))^{-\frac{1}{2}}$ and $b = \xi(4(1 - e^{-(\tau-s)})^{\frac{1}{2}})$ and (5.27) to show that

$$|J| \leq C\int_{\tau-1}^{\tau}(1 - e^{-(\tau-s)})^{-1}\int_{|\xi|\leq\varepsilon(s)}(1 - e^{-(\tau-s)})^{-1}w(y,\tau - s;0)|y||\xi|d\xi$$

$$\leq C\int_{\tau-1}^{\tau}(1 - e^{-s})^2\varepsilon(s)^3|y|\exp\left(-\frac{y^2e^{-s}}{(1 - e^{-s})}\right)ds \leq C|y|\int_0^1\varepsilon(\tau - s)^3 s^{-2}e^{-\frac{Cy^2}{s}}ds. \tag{5.29}$$

Putting together (5.25) and (5.29), the result follows. □

For latter reference, we summarize the results obtained in Lemmas 5.2–5.4 as follows.

COROLLARY 5.5. *Let $E(y,\tau)$, $Q(y,\tau)$ be functions such that* (i) $E(y,\tau_0) = Q(y,\tau_0)$ *and* (ii) E *and* Q *solve, respectively,* (5.4) *and* (5.5) *for* $\tau > \tau_0$. *Assume also that* (4.2) *holds. Then for any $R > 0$, there exists $C > 0$ such that*

$$|E(y,\tau) - Q(y,\tau)| \leq C\left(\frac{\varepsilon(\tau)^3}{y}\right) \tag{5.30}$$

whenever $\bar{\varepsilon}(\tau_0)^{\frac{1}{4}} \leq y \leq R$ *and* $\tau > \tau_0$

Proof. The proof follows from (5.17), (5.23), and the bounds (4.2). □

6. Derivation of (1.2): Analysis of the inner region.

Let $\sigma(\tau)$ be a function to be discussed later (cf. (6.9)). We now fix $\bar{\tau} \gg 1$ and define

$$\xi = \frac{y}{\sigma(\bar{\tau})}; \qquad w(\xi,\tau) = (\sigma(\bar{\tau}))^{-2}\Phi(\sigma(\bar{\tau})\xi,\tau), \tag{6.1}$$

where $\Phi(y,\tau)$ is given in (2.3). A quick computation reveals that $w(\xi,\tau)$ satisfies

$$(\sigma(\bar{\tau}))^2 w_\tau = \Delta w - H(w) + (\sigma(\bar{\tau}))^2\left(w - \frac{\xi\nabla w}{2}\right), \tag{6.2}$$

where the operators Δ and ∇ are now written with respect to the inner space variable ξ. When determining the asymptotics of solutions of (6.2), a key role is played by the stationary equation

$$\Delta\nu = H(\nu). \tag{6.3}$$

For any $\lambda > 0$, a radial solution of (6.3) is given by $\nu_\lambda(\xi) = \lambda^2\bar{\nu}(\frac{\xi}{\lambda})$, where

$$\bar{\nu}(r) = \frac{r^2}{4} - \frac{1}{4} - \frac{1}{2}\log r. \tag{6.4}$$

We can readily check that the radial, nontrivial solution of (6.3) which satisfies $\nu(\xi) = 0$ for $\xi \leq \lambda$ and $\nu(\lambda) = \nu'(\lambda) = 0$ is given by

$$\nu_\lambda(\xi) = \frac{\xi^2}{4} - \frac{\lambda^2}{2}\log\left(\frac{\xi}{\lambda}\right) - \frac{\lambda^2}{4} \quad \text{for } \xi > \lambda. \tag{6.5}$$

It will be convenient to compare the functions $\nu_\lambda(\xi)$ given in (6.5) with the stationary solution of (6.2) that takes off at $\xi = \lambda$. The corresponding result reads as follows.

Lemma 6.1. *Let* $\tilde{w}_\lambda(\xi) \equiv \tilde{w}_\lambda(\xi;\bar{\tau})$ *be the stationary solution of* (6.2) *such that* $\tilde{w}_\lambda(\lambda) = \tilde{w}'_\lambda(\lambda) = 0$ *and* $\tilde{w}_\lambda(\xi) > 0$ *for* $\xi > \lambda > 0$. *Then there holds*

$$\tilde{w}_\lambda(\xi) = \nu_\lambda(\xi) + O(\sigma^2\lambda^2 \log\lambda) \quad \textit{for } \xi \le 1. \tag{6.6}$$

Proof. We set $\tilde{w}_\lambda = \nu_\lambda + \varphi$. A quick check reveals that φ solves

$$\varphi'' + \frac{\varphi'}{\xi} + \sigma^2\left(\varphi - \frac{\xi\varphi'}{2} - \frac{\lambda^2}{2}\log\frac{\xi}{\lambda} - \frac{\lambda^2}{4}\right) = 0,$$
$$\varphi(\lambda) = \varphi'(\lambda) = 0.$$

Consider first the case where ξ is close to λ. Standard ordinary differential equation (ODE) arguments yield that, in such a region

$$\varphi(\xi) \sim C\sigma^2\lambda^2\xi^2 \log\frac{\xi}{\lambda} \quad \text{for some real } C. \tag{6.7}$$

When $\lambda \ll \xi \le 1$, we introduce a new variable $\eta = \frac{\xi}{\lambda}$. Setting $\dot{\varphi} = \frac{d\varphi}{d\eta}$, we readily check that φ satisfies

$$\ddot{\varphi} + \frac{\dot{\varphi}}{\eta} + \sigma^2\lambda^2\left(\varphi - \frac{\eta\dot{\varphi}}{2}\right) = \frac{\sigma^2\lambda^4}{2}\log\eta + \frac{\sigma^2\lambda^2}{4}.$$

A dominated balance argument shows that the third term on the left is negligible with respect to the remaining ones. This in turn implies that $\varphi(\eta) \sim K^2\sigma^2\lambda^4\eta^2\log\eta$. Back to the original variables, we have derived

$$\varphi(\xi) \sim K\sigma^2\lambda^2\xi^2 \log\frac{\xi}{\lambda} \quad \text{for } \lambda \ll \xi \le 1. \tag{6.8}$$

Matching (6.7) and (6.8), we obtain $C = K$ and (6.6) follows. □

Let us now define

$$\sigma(\tau) = (\varepsilon(\tau))^\theta, \quad \text{where } \theta \text{ is a positive and small number}, 0 < \theta < \frac{1}{4}, \tag{6.9}$$

$$W(\tau) = \frac{1}{4} + \frac{1}{(\sigma(\bar{\tau}))^2}\left(a_0(\tau)\varphi_0 + a_1(\tau)\varphi_1(\sigma(\bar{\tau})) + E(\sigma(\bar{\tau}),\tau)\right), \tag{6.10}$$

where a_0, a_1, and E are as in (5.3), and

$$\lambda(\tau) = \frac{\varepsilon(\tau)}{\sigma(\tau)}. \tag{6.11}$$

We shall prove the following lemma.

Lemma 6.2. *Assume that conditions* (4.2) *and* (4.11) *hold. Then there exists a constant* $C > 0$ *such that*

$$|W(\tau) - W(\bar{\tau})| \le \frac{C}{\tau}\left(\frac{\varepsilon(\bar{\tau})}{\sigma(\bar{\tau})}\right)^2, \tag{6.12}$$

provided that $|\tau - \bar{\tau}| \le \frac{1}{\tau}$ *and* $\bar{\tau} \ge \tau_0 \gg 1$.

Proof. We set

$$\begin{aligned} W(\tau)-W(\bar\tau) &= \varphi_0(\sigma(\bar\tau))^{-2}(a_0(\tau)-a_0(\bar\tau))+\varphi_1(\sigma(\bar\tau))^{-2}(a_1(\tau)-a_1(\bar\tau)) \\ &\quad + (\sigma(\bar\tau))^{-2}(E(\sigma(\bar\tau),\tau)-E(\sigma(\bar\tau),\bar\tau)) \equiv W_1+W_2+W_3. \end{aligned} \tag{6.13}$$

Terms W_1 and W_2 in (6.13) are easily dealt with. For instance, since $\psi(y,\tau)$ satisfies (5.1), one sees that if $\tau>\bar\tau$,

$$\begin{aligned} W_2 &= \varphi_1(\sigma(\bar\tau))(\sigma(\bar\tau))^{-2}\int_{\bar\tau}^{\tau}\langle\chi_\varepsilon(s),\varphi_1\rangle ds \\ &= \varphi_1(\sigma(\bar\tau))(\sigma(\bar\tau))^{-2}\int_{\bar\tau}^{\tau}\int_{|y|\le\varepsilon(s)}\varphi_1(y)e^{-\frac{y^2}{4}}dyds; \end{aligned}$$

hence

$$|W_2|\le\frac{C|\tau-\bar\tau|}{(\sigma(\bar\tau))^2}((\varepsilon(s)-\varepsilon(\bar\tau))^2+\varepsilon(\bar\tau)^2)\le\frac{C}{\tau}\left(\frac{\varepsilon(\bar\tau)}{\sigma(\bar\tau)}\right)^2\left(1+\frac{M^2}{\tau^3}\right),$$

where (4.2) has been used to obtain the last inequality above. A similar bound for W_1 is obtained by means of (4.11). To estimate W_3, we first observe that

$$|Q(\sigma(\bar\tau),\tau)-Q(\sigma(\bar\tau),\bar\tau)|\le C\frac{\bar\varepsilon(\tau)^2}{\tau}\quad\text{for } |\tau-\bar\tau|\le\frac{1}{\tau}. \tag{6.14}$$

To obtain (6.14), we consider first the case where $|\tau_0-\bar\tau|\le\frac{1}{\tau_0}$. Setting

$$D(y)=\delta(y)-\sum_{k=0}^{1}\langle\varphi_k,\delta(y)\rangle\varphi_k(y)$$

it then turns out that

$$\begin{aligned} Q(y,\tau) &= \gamma\varepsilon(\tau_0)^2S(\tau-\tau_0)F(y)+\gamma\int_{\tau_0}^{\tau}S(\tau-s)(\varepsilon(s)^2D)ds \\ &= \gamma\varepsilon(\tau_0)^2S(\tau-\tau_0)F(y)+\gamma\int_{\tau_0}^{\tau}S(\tau-s)(\varepsilon(\tau_0)^2D)ds \\ &\quad+\gamma\int_{\tau_0}^{\tau}S(\tau-s)((\varepsilon(s)^2-\varepsilon(\tau_0)^2)D)ds \\ &= \gamma\varepsilon(\tau_0)^2F(y)+O(\varepsilon(\tau)^2\tau^{-\frac{3}{2}}|\log y|) \end{aligned}$$

for $|y|$ small, whereupon (6.14) follows. On the other hand, by (5.30) we have that

$$|E(\sigma(\bar\tau),\tau)-E(\sigma(\bar\tau),\bar\tau)|\le C\frac{\varepsilon(\tau)^2}{\tau}|Q(\sigma(\bar\tau,\tau))-Q(\sigma(\bar\tau),\bar\tau)|. \tag{6.15}$$

We thus obtain from (6.14) and (6.15) that

$$|W_3|\le C\left(\frac{\varepsilon(\tau)^2}{\tau}\right)\quad\text{for } |\tau_0-\bar\tau|\le\frac{1}{\tau_0}\ \text{ and }\ \bar\tau\ge\tau_0\gg1,$$

and putting together the bounds obtained for W_i $(i = 1, 2, 3)$ the proof is concluded in this case.

When $|\tau_0 - \bar{\tau}| > \frac{1}{\tau_0}$, we make use of (5.12) to check that (6.14) continues to hold. This is done by comparing the different terms appearing in the right-hand side of (5.12) when evaluated at $(\sigma(\bar{\tau}), \tau)$ and $(\sigma(\bar{\tau}), \bar{\tau})$, respectively. A typical argument in this direction goes as follows. For $i = 1, 2$ set $\sum_i = (\delta(\bar{\tau}))^2(4(e^{\beta_i} - 1))^{-1}$, where $\beta_i = \max\{\tau_i - 1, \tau_0\}$, and let us write $s_i = \tau_i - \log(1 - (\delta(\bar{\tau}))^2/4\eta)$. Then, if $|s_1 - s_2| \leq \frac{1}{\bar{\tau}}$, we have that

$$\begin{aligned}
&\left| \int_{\Sigma_1}^{\infty} \varepsilon(s_1)^2 e^{-\eta}\eta^{-1} d\eta - \textstyle\int_{\Sigma_2}^{\infty} \varepsilon(s_2)^2 e^{-\eta}\eta^{-1} d\eta \right| \\
&\quad \leq \int_{\Sigma_1}^{\infty} |\varepsilon(s_1) + \varepsilon(s_2)||\varepsilon(s_1) - \varepsilon(s_2)| e^{-\eta}\eta^{-1} d\eta + \int_{\Sigma_1}^{\Sigma_2} \varepsilon(s_2)^2 e^{-\eta}\eta^{-1} d\eta \\
&\quad \leq \ CM(\bar{\varepsilon}(\tau))^2 \tau^{-\frac{3}{2}} |\log \Sigma_1| + C(\bar{\varepsilon}(\tau))^2(\delta(\bar{\tau}))^2 |\log \Sigma_1| \\
&\quad \leq \ CM\theta \frac{(\bar{\varepsilon}(\tau))^2}{\tau},
\end{aligned}$$

where $\theta > 0$ can be selected arbitrarily small as $\tau_0 \to \infty$, and we have assumed for definiteness that $\Sigma_1 < \Sigma_2$. We omit further details. □

A key result in this section is the following.

LEMMA 6.3. *Under the assumptions of Lemma* 6.2*, there exists a constant* $C > 0$ *such that*

$$|W(\tau) - \nu_{\lambda(\tau)}(1)| \leq \frac{C}{\tau} \left(\frac{\varepsilon(\bar{\tau})}{\sigma(\bar{\tau})} \right)^2 \tag{6.16}$$

provided that $|\tau - \bar{\tau}| \leq \frac{1}{\tau}$*, where* ν_λ *is given in* (6.5) *and* $\bar{\tau} \geq \tau_0 \gg 1$.

Proof. We shall argue by contradiction and therefore assume that for any $K > 0$ there exists $\bar{\tau} \gg 1$ and $\tilde{\tau} \in (\bar{\tau} - \frac{1}{\tau}, \bar{\tau} + \frac{1}{\bar{\tau}})$ such that

$$|W(\tilde{\tau}) - \nu_{\lambda(\tilde{\tau})}(1)| > \frac{K}{\tilde{\tau}} \left(\frac{\varepsilon(\bar{\tau})}{\sigma(\bar{\tau})} \right)^2. \tag{6.17}$$

Now let τ be any time in the interval $(\bar{\tau} - \frac{1}{\bar{\bar{\tau}}}, \bar{\tau} + \frac{1}{\tilde{\bar{\tau}}})$. In view of (6.12) and (6.17), it holds that

$$|W(\tau) - \nu_{\lambda(\tilde{\tau})}(1)| > \frac{K}{2\tilde{\tau}} \left(\frac{\varepsilon(\tilde{\tau})}{\sigma(\tilde{\tau})} \right)^2. \tag{6.18a}$$

Assume for definiteness that

$$|W(\tau) - \nu_{\lambda(\tilde{\tau})}(1)| = \upsilon_{\lambda(\tilde{\tau})}(1) - W(\tau). \tag{6.18b}$$

We now claim the following:

$$\begin{aligned} &\text{There exists } \mu > 0 \quad \text{such that if} \quad \lambda_0(\tilde{\tau}) = \lambda(\tilde{\tau})(1 + \mu(\tilde{\tau})^{-\frac{3}{2}}), \\ &\qquad\qquad\qquad\qquad \text{then} \quad \nu_{\lambda_0}(1) > W(\tau). \end{aligned} \tag{6.19}$$

To check (6.19), we observe that since $\tilde{\tau} \gg 1$,

$$\begin{aligned}
\nu_{\lambda_0}(1) &= \frac{1}{4} + \frac{\lambda_0^2}{2}\log\lambda_0 - \frac{\lambda_0^2}{4} \\
&\sim \frac{1}{4} + \frac{\lambda^2}{2}\left(1+\mu(\tilde{\tau})^{-\frac{3}{2}}\right)^2 \log\lambda_0 - \frac{\lambda_0^2}{4} \\
&\sim \frac{1}{4} + \frac{\lambda(\tilde{\tau})^2}{2}\left(1+2\mu(\tilde{\tau})^{-\frac{3}{2}}\right)\left(\log\lambda(\tilde{\tau}) + \log\left(1+\mu(\tilde{\tau})^{-\frac{3}{2}}\right)\right) \\
&\quad - \frac{\lambda(\tilde{\tau})^2}{4}\left(1+2\mu(\tilde{\tau})^{-\frac{3}{2}}\right) = \frac{1}{4} + \frac{\lambda(\tilde{\tau})^2}{2}\log\lambda(\tilde{\tau}) \\
&\quad + \mu\lambda(\tilde{\tau})^2(\tilde{\tau})^{-\frac{3}{2}}\log\lambda(\tilde{\tau}) - \frac{\lambda(\tilde{\tau})^2}{4} \qquad (6.20)
\end{aligned}$$

whereas by (6.18),

$$W(\tau) < \nu_{\lambda(\tilde{\tau})}(1) - \frac{K}{2\tilde{\tau}}\left(\frac{\varepsilon(\tilde{\tau})}{\sigma(\tilde{\tau})}\right)^2 = \frac{1}{4} + \frac{\lambda(\tilde{\tau})^2}{2}\log\lambda(\tilde{\tau}) - \frac{\lambda(\tilde{\tau})^2}{4} - \frac{K}{2\tilde{\tau}}\lambda(\tilde{\tau})^2. \tag{6.21}$$

From (6.20) and (6.21), it follows that (6.19) holds provided that

$$\frac{K}{2\tilde{\tau}} > \frac{\mu}{\tilde{\tau}^{\frac{3}{2}}}|\log\lambda(\tilde{\tau})|,$$

and this last inequality is satisfied by selecting $\mu > 0$ small enough since

$$\tau^{-\frac{3}{2}}\log\lambda(\tau) \sim (1-\theta)\tau^{-1} \quad \text{as } \tau \to \infty.$$

We now set

$$z(\xi,\tau) = (w(\xi,\tau) - \tilde{w}_{\lambda_0}(\xi))_+, \quad \text{where } s_+ = \max\{s,0\}.$$

Since $(H(s)-H(t))(s-t)^{-1} \geq 0$ whenever $s \neq t$, we can readily check that z satisfies

$$z_\tau \leq (\sigma(\tilde{\tau}))^{-2}\Delta z + \left(z - \frac{\xi\nabla z}{2}\right) \quad \text{for } \tau > \tilde{\tau} - \frac{1}{\tilde{\tau}}, \quad 0 < \xi < 1, \tag{6.22a}$$

whereas by (6.19),

$$z = 0 \quad \text{when } \xi = 0, 1 \text{ and } \tau > \tilde{\tau} - \frac{1}{\tilde{\tau}} \tag{6.22b}$$

and

$$z = O\left(\frac{1}{\sigma(\tilde{\tau})^2}\right) \quad \text{at } \tau = \tilde{\tau} - \frac{1}{\tilde{\tau}}. \tag{6.22c}$$

By classic parabolic theory, it follows from (6.22) that

$$z(\xi,\tau) \leq A(\sigma(\tilde{\tau}))^{-2}\exp\left(-\frac{A(\tau-\tilde{\tau})}{(\sigma(\tilde{\tau})^2}\right) \quad \text{in } Q,$$

where

$$Q = \left\{ (\xi, \tau) : |\xi| \le 1, \quad \tau \in \left(\bar{\tau} - \frac{1}{\tilde{\tau}}, \tilde{\tau} + \frac{1}{\tilde{\tau}} \right) \right\}.$$

In particular, $w(\xi,\tau) \sim \tilde{w}_{\lambda_0}(\xi)$ at $\tau = \tilde{\tau}$. Recalling (6.6) (with σ, λ replaced by $\sigma(\tilde{\tau})$, $\lambda_0(\tilde{\tau})$, respectively), we see that the (rescaled) free boundary of $w(\xi,\tilde{\tau})$ is very close to $\lambda_0(\tilde{\tau})$ and in particular is larger than $\lambda(\tilde{\tau})$, which is a contradiction. The case where (6.18b) is replaced by $|W(\tau) - \nu_{\lambda(\tilde{\tau})}(1)| = W(\tau) - \nu_{\lambda(\tilde{\tau})}(1)$ is similar and will be omitted. □

We now point out the following consequence of Lemmas 6.2 and 6.3.

Corollary 6.4. *There holds*

$$|\varepsilon(\tau) - \varepsilon(\tilde{\tau})| \le C\tau^{-\frac{3}{2}}\varepsilon(\tilde{\tau}) \quad \textit{for some } C > 0 \tag{6.23}$$

whenever $|\tau - \tilde{\tau}| < \frac{1}{\tau}$ *and* $\tilde{\tau} \ge \tau_0 \gg 1$.

Proof. From (6.12) and (6.16) we readily see that

$$|\nu_{\lambda(\tau)}(1) - \nu_{\lambda(\tilde{\tau})}(1)| \le \frac{2C}{\tau}(\lambda(\tilde{\tau}))^2 \quad \text{for } |\tau - \tilde{\tau}| < \frac{1}{\tau}$$

and the result follows at once in view of the explicit formula (6.5). □

We shall conclude the proof of (1.2) by means of a careful analysis of (6.16), which can be thought of as an integral equation for the unknown rescaled free boundary $\varepsilon(\tau)$. Assume now that (4.2) holds. Then in view of (5.3), (5.30), and (6.16), we have that

$$\left| a_0(\tau)\varphi_0 + a_1(\tau)\varphi_1(0) + Q(y,\tau) - \frac{\varepsilon^2(\tau)}{2}\log\left(\frac{\varepsilon(\tau)}{\sigma(\tilde{\tau})}\right) + \frac{\varepsilon(\tau)^2}{4} \right| = O\left(\frac{\varepsilon(\tau)^2}{\tau}\right). \tag{6.24}$$

Note that the error involved in replacing $\varphi_1(y)$ by $\varphi_1(0)$ is already accounted for in the right-hand side of (6.24). We now take advantage of (5.12) to estimate $Q(y,\tau)$ in (6.24). Recalling (6.23), we have that for $(z(\tau))^0 \le y \le 1$ and Σ as in (5.12),

$$\begin{aligned} &\left| \int_\Sigma^\infty \varepsilon(s)^2 e^{-\eta}\eta^{-1}d\eta - \varepsilon(\tau)^2 \int_\Sigma^\infty e^{-\eta}\eta^{-1}d\eta \right| \\ &\le C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}} \int_\Sigma^\infty e^{-\eta}\eta^{-1}d\eta \\ &\le C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}}(1 - |\log\Sigma|)e^{-\Sigma} \le C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}}. \end{aligned} \tag{6.25a}$$

Notice that (6.23) provides a factor $\tau^{-\frac{3}{2}}$ in the right-hand side of the first inequality above in sets where $|s - \tau| < \frac{1}{\tau}$. Extending such a bound to the interval $|s - \tau| \le 1$ required by our choice of Σ yields then the final factor $\tau^{-\frac{1}{2}}$. A similar argument gives

$$\left| \int_\Sigma^\infty \varepsilon(s)^2\eta^2 e^{-\eta}(4\eta^2)^{-1}d\eta - \varepsilon(\tau)^2 \int_\Sigma^\infty \eta^2 e^{-\eta}(4\eta^2)^{-1}d\eta \right| \le C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}} \tag{6.25b}$$

and

$$\begin{aligned} &\left| \int_{\tau-1}^\tau (A_1 + A_2(1 - e^{-(\tau-s)}))e^{\tau-s}\varepsilon(s)^2 ds \right. \\ &\left. \quad - (\varepsilon(\tau))^2 \int_{\tau-1}^\tau (A_1 + A_2(1 - e^{-(\tau-1)}))e^{\tau-s}ds \right| \le C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}}. \end{aligned} \tag{6.25c}$$

Recalling (5.12), we have obtained the following estimate:

$$\tag{6.26} \begin{aligned} &\left| Q(y,\tau) - \gamma - \varepsilon(\tau_0)^2 S(\tau-\tau_0)F(y) - \gamma\varepsilon(\tau_0)^2 \int_{\tau_0}^{\tau} S(\tau-s)\left(\delta(y) - \sum_{k=0}^{1}\langle\varphi_k, \delta(y)\rangle\varphi_k\right) ds \right| \\ &\leq C(\bar{\varepsilon}(\tau))^2 \tau^{-\frac{1}{2}}. \end{aligned}$$

Keeping in mind the definition of $F(y)$ (cf. (3.8)), we see that

$$S(\tau-\tau_0)F(y) + \int_{\tau_0}^{\tau} S(\tau-s)\left(\delta(y) - \sum_{k=0}^{1}\langle\varphi_k, \delta(y)\rangle\varphi_k\right) ds = F(y).$$

It then turns out that (6.26) can be recast in the form

$$\begin{aligned} &|Q(y,\tau) - \gamma\varepsilon(\tau)^2 F(y) - \gamma(\varepsilon(\tilde{\tau})^2 - \varepsilon(\tau)^2) S(\tau-\tau_0)F(y)| \\ &\leq C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}} \quad \text{for } (\bar{\varepsilon}(\tau))^\theta \leq y \leq 1. \end{aligned}$$

Since the term $S(\tau-\tau_0)F(y)$ decays exponentially on sets where $|y|$ is bounded, we may take advantage again of (6.23) to obtain that

$$\tag{6.27} |Q(y,\tau) - \gamma\varepsilon(\tau)^2 F(y)| \leq C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}} \quad \text{whenever } (\bar{\varepsilon}(\tau))^\theta \leq y \leq 1.$$

Using now the explicit representation for $F(y)$ (cf. (3.9)), we deduce from (6.24) and (6.27) that

$$\tag{6.28} \left| a_0(\tau)\varphi_0 + a_1(\tau)\varphi_1(0) + \left(B\gamma + \frac{1}{4}\right)\varepsilon(\tau)^2 - \frac{\varepsilon(\tau)^2}{2}\log\varepsilon(\tau) \right| \leq C\bar{\varepsilon}(\tau)^2\tau^{-\frac{1}{2}}.$$

Let us now define $\varepsilon^*(\tau)$ as follows

$$\varepsilon^*(\tau) = \begin{cases} \varepsilon(\tau) & \text{if } \tau_0 \leq \tau \leq \tau_1, \\ \bar{\varepsilon}(\tau) & \text{if } \tau \geq \tau_1. \end{cases}$$

We next observe that if (4.11) is satisfied there holds

$$a_k(\tau) = -\int_{\tau}^{\infty} e^{\lambda_k(\tau-s)}\langle\chi_{\varepsilon^*(\tau)}, \varphi_k\rangle ds.$$

Substituting this into (6.28), we finally arrive at

$$\tag{6.29} \begin{aligned} &\left| \left(B\gamma + \frac{1}{4}\right)\varepsilon^*(\tau)^2 - \frac{\varepsilon^*(\tau)^2}{2}\log\varepsilon^*(\tau) \right. \\ &\left. - \sum_{k=0}^{1}\varphi_k(0)\int_{\tau}^{\infty} e^{(1-k)(\tau-s)}\langle\chi_{\varepsilon^*(s)}, \varphi_k\rangle ds \right| \\ &= 0(\varepsilon^*(\tau)^2\tau^{-1/2}) \quad \text{for } \tau \gg 1. \end{aligned}$$

This is essentially the integral equation that has been studied in detail in section 3 (cf. the argument following (3.17)). In view of our previous analysis in section 3, we may summarize our discussion in the following.

LEMMA 6.5. *Assume that* (4.2) *and* (4.11) *hold. We then have that*

$$\varepsilon(\tau) = \bar{\varepsilon}(\tau)(1 + o(1)) \quad \textit{for } \tau \gg 1 \tag{6.30}$$

uniformly on $\tau_0 \le \tau \le \tau_1$.

End of the proof for Proposition 4.1. Having obtained (6.30) under assumption (4.11), it merely remains to show that condition $|\psi(y,\tau)| < M(1+y^2)$ in (4.8) holds for some constant M which is independent of the size of the interval $[\tau_0, \tau_1]$. To check this point, we argue as follows. We have just seen that

$$\psi(y,\tau) \sim a_0(\tau)\varphi_0 + a_1(\tau)\varphi_1(y) + O(\varepsilon(\tau)^2) \quad \text{for } \tau \gg 1 \text{ and } y = O(1). \tag{6.31}$$

We now claim that we may formally differentiate twice with respect to y in both sides of (6.31) and the corresponding expansion still holds. To wit, we set $z(y,\tau) = \psi(y,\tau) - a_1(\tau)\varphi_1(y)$ and remark that in regions where $y = O(1)$, one has that $z = O(\varepsilon^2(\tau))$ and satisfies

$$Lz = 0(\varepsilon^2(\tau)\varphi_1(y)),$$

where L denotes the parabolic operator in (5.1). It then follows that for $\tau \ge \tau_0 \gg 1$

$$z(y,\tau) = \exp\left(\int_{\tau_0}^{\tau} D(s)ds\right)\varphi_1(y) \quad \text{with } D(s) = O(\varepsilon(s)^2)$$

whereupon the desired bound for ψ follows. □

Proof of (1.6) *in Theorem* 1.3. It has been shown above that

$$\psi_{yy}(y,\tau) \sim Ca_1(\tau) + O(\varepsilon^2(\tau)) \quad \text{with } C = \frac{d^2}{dy^2}(\varphi_1(y)) \tag{6.32}$$

for, say, $y = O(1)$ and $\tau \gg 1$. Since $\theta(r,t) = \psi_{yy} + \frac{\psi_y}{y} = Ca_1(\tau)$ for some $C_1 > 0$, it follows that, setting $r = Ae^{-\tau/2}$ with $A > 0$,

$$\theta(r,T) \sim C_1 a_1 \left(-2\log\frac{r}{A}\right)(1 + o(1)) \quad \text{as } \tau \to 0. \tag{6.33}$$

Since

$$a_1(\tau) \sim \frac{\varepsilon^2(\tau)\log\varepsilon(\tau)}{2} \quad \text{as } \tau \to \infty,$$

the result follows at once from (6.33) and (3.18). □

$$f(x,T) \sim Ca_1\left(-2\log\frac{x}{A}\right)(1 + o(1)) \quad \text{as } x \to 0. \tag{6.34}$$

Since $a_1(\tau) \sim \frac{1}{2}\varepsilon^2(\tau)\log\varepsilon(\tau)$ as $\tau \to \infty$, the result follows at once from (6.34) and (3.18). □

7. The remaining cases. In this final section we shall merely sketch those modifications of the arguments developed in sections 4–6 which are required to obtain (1.3)–(1.5) and (1.7)–(1.9), thus concluding the proofs of Theorems 1.1–1.3.

7.1. Obtaining (1.3) in Theorem 1.1. In the topological argument described in section 4, we must replace $\bar{\varepsilon}(\tau)$ in (4.1) by

$$\bar{\varepsilon}(\tau) = Ce^{(1-\frac{l}{2})\tau}\tau^{-1/l-1},$$

where C is an arbitrary constant and l is an integer such that $l \geq 2$. We then substitute (4.2a) by

$$\sup\left\{|\varepsilon(\tau) - \bar{\varepsilon}(s)|, \text{where } \tau, s \in [\tau_0, \tau_1] \text{ and } |\tau - s| < \frac{1}{\tau}\right\} < M\bar{\varepsilon}(\tau)\tau^{-1},$$

which can be rephrased in an informal way as requiring that $|\frac{d}{d\tau}\bar{\varepsilon}(\tau)| < M\bar{\varepsilon}(\tau)$. Condition (4.2b) is then kept as before. As for (4.11), it is to be replaced by

$$a_k(\tau) = -\int_\tau^\infty e^{(1-\frac{k}{2})(\tau-s)}\langle \chi_{\bar{\varepsilon}(s)}, \varphi_k\rangle ds$$

for $k = 0, 1, 2, \ldots, l$.

With these modifications in mind, the analogue of Proposition 4.1 is readily stated. To analyze the outer region in this case, one writes $\psi(y, \tau)$ in the form

$$\psi(y,\tau) = \sum_{k=0}^{l-1} a_k(\tau)\varphi_k(y) + a_l(\tau)\varphi_l(y) + E(y,\tau). \tag{7.1}$$

The term $E(y, \tau)$ in (7.1) will be approximated as before by $Q(y, \tau)$, where Q satisfies now (3.25b) instead of (5.5). The solution of such an equation for $\tau > \tau_0$ such that $Q(y, \tau_0) = 0$ is given by

$$Q_0(y,\tau) = \gamma\int_{\tau_0}^{\tau} K_l(y, \tau - s)e^{\tau-s}\varepsilon(s)^2 ds, \tag{7.2a}$$

where

$$K_l(y,\tau) = (4\pi(1-e^{-\tau}))^{-1}\left(\exp\left(-\frac{y^2e^{-\tau}}{4(1-e^{-\tau})}\right) - \sum_{k=0}^{l}\varphi_k\left\langle\varphi_k, \exp\left(-\frac{y^2e^{-\tau}}{4(1-e^{-\tau})}\right)\right\rangle\right). \tag{7.2b}$$

Notice that $|K_l(y, \tau)| \leq Ce^{-l\tau}$ uniformly on sets $|y| \leq R < \infty$ when $\tau \geq 1$. Arguing as for Lemma 5.2, we then obtain the corresponding version of that result in our case. This last is obtained by making a few modifications in (5.12):

(a) Replace $S(\tau)$ by $S_l(\tau)$ there, where $S_l(\tau)$ is the semigroup associated to operator A_l in (3.27c).

(b) Replace $F(y)$ given in (3.8) by the corresponding solution of (3.27).

(c) Substitute $K(y, \tau)$ by $K_l(y, \tau)$ given in (7.2b).

(d) Replace the factor $(A_1 + A_2(1 - e^{-(\tau-1)}))$ in the integral where it appears in (5.12) by $(A_1 + A_2(1 - e^{-(\tau-1)}) + \cdots + A_l(1 - e^{-(\tau-s)}))$.

The analysis of the inner region is then performed as in section 6. The integral equation (6.16) is now to be replaced by

$$|W(\tau) - \nu_{\lambda(\tau)}(1)| \leq C\varepsilon(\tau)^2,$$

provided that $|\tau - \bar{\tau}| \leq \frac{1}{\tau}, \bar{\tau} \gg 1$.

Arguing as in section 6, we then arrive at an integral equation which is similar to (3.30). We shall omit further details.

7.2. The case where $N \geq 3$: End of the proof of Theorem 1.2. We first describe the main steps toward obtaining (1.4). To begin with, we replace $\bar{\varepsilon}(\tau)$ in (4.1) by

$$\bar{\varepsilon}(\tau) = C\tau^{-\frac{1}{N-2}}, \quad C > 0,$$

and replace (4.2a) by

$$|\varepsilon(\tau) - \bar{\varepsilon}(s)| < M\bar{\varepsilon}(s)\tau^{-(1-\theta)} \quad \text{for some } 0 < \theta \ll 1$$

whenever $|\tau - s| \leq \tau^{\theta}$.

The main novelty now with respect to the previous cases is that we may estimate directly the error term $E(y,\tau)$ in (5.3) to obtain

$$|E(y,\tau)| \leq C\bar{\varepsilon}(\tau)^N(1 + y^{-(N-2)}) \quad \text{for } \tau \gg 1$$

uniformly on sets $y \leq R < \infty$. We are thus led to a version of the integral equation (6.16) which reads now as follows:

$$|W(\tau) - \nu_{\lambda(\tau)}(1)| \leq C\bar{\varepsilon}(\tau)^2\tau^{-(1-\theta)} \quad \text{for } |\tau - \bar{\tau}| \leq \tau^{\theta}, \bar{\tau} \gg 1,$$

whence the statement in Lemma 6.5 follows in this case.

Finally, (1.5) corresponds to the situation where $\chi_{\varepsilon(\tau)}$ can be asymptotically neglected in (5.1), so that the analysis sketched at the end of section 3 can be carried out in a straightforward way.

7.3. End of the proof of Theorem 1.3. To obtain (1.7), we can argue as in the last part of section 6 to obtain that

$$\psi(y,\tau) \sim a_l(\tau)\varphi_l(y) \quad \text{for } y = O(1) \text{ and } \tau \gg 1,$$

where the expansion above also holds when differentiated twice with respect to y. In view of (6.33), one then has that at points $x = Ae^{-\tau/2}$,

$$\theta(x,T) \sim a_l\left(-2\log\frac{x}{A}\right)A^{2l-2} = a_l\left(-2\log\frac{x}{A}\right)|x|^{2l-2}e^{(l-1)\tau},$$

whence (1.7) follows in view of (3.30) and (3.31).

As to (1.8), we make use of (3.33) and (3.34) to observe that

$$\psi_{yy}(y,\tau) \sim C\tau^{-\frac{2}{N-2}} \quad \text{for } y = O(1) \text{ and } \tau \gg 1.$$

This readily gives that for $x = Ae^{-\tau/2}$

$$\theta(x,T) \sim C(|\log|x||)^{-\frac{2}{N-2}} \quad \text{as } x \downarrow 0.$$

Finally, to obtain (1.9) we merely recall that, in view of (3.35), one has that for $x = Ae^{-\tau/2}$

$$\psi_{yy}(A,\tau) \sim e^{(1-l)\tau}A^{2l-2},$$

whereupon $\theta(x,T)$ is shown to be such that

$$\theta(x,T) \sim C|x|^{2l-2} \quad \text{as } x \downarrow 0.$$

REFERENCES

[AK] D. G. Aronson and S. Kamin, *Disappearance of phase in the Stefan problem: One space dimension*, preprint 614, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, 1990.

[AV] S. B. Angenent and J. J. L. Velázquez, *Degenerate neckpinches in mean curvature flow*, J. Reine Angew. Math., to appear.

[B1] A. Bressan, *On the asymptotic shape of blow up*, Indiana University Math. J., 39 (1990), pp. 947–959.

[B2] A. Bressan, *Stable blow up patterns*, J. Differential Equations, 98 (1992), pp. 57–75.

[DH] G. B. Davies and J. M. Hill, *A moving boundary problem for the sphere*, IMA J. Appl. Math., 29 (1982), pp. 99–111.

[EK] L. C. Evans and B. F. Knerr, *Instantaneous shrinking of the support of nonnegative solutions to certain nonlinear parabolic equations and variational inequalities*, Illinois J. Math., 23 (1979), pp. 153–166.

[GK] Y. Giga and R. V. Kohn, *Asymptotically self-similar blow up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[HD1] J. M. Hill and J. Dewynne, *On an integral formulation for moving boundary problems*, Quart. Appl. Math., XLI (1984), pp. 443–455.

[HD2] J. M. Hill and J. Dewynne, *On the inward solidification of cylinders*, Quart. Appl. Math., XLIV (1986), pp. 59–70.

[HV1] M. A. Herrero and J. J. L. Velázquez, *Approaching an extinction point in one-dimensional semilinear heat equations with strong absorption*, J. Math. Anal. Appl., 170 (1992), pp. 353–381.

[HV2] M. A. Herrero and J. J. L. Velázquez, *Singularity formulation in the one-dimensional supercooled Stefan problem*, European J. Appl. Math., 7 (1996), pp. 119–150.

[HV3] M. A. Herrero and J. J. L. Velázquez, *Explosion de solutions d'equations paraboliques semilinéaires supercritiques*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 141–143.

[L] N. N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.

[M] A. M. Meirmanov, *The Stefan Problem*, Expositions in Mathematics Series, De Gruyter, Hawthorne, NY, 1992.

[MF] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vol. I, McGraw–Hill, New York, 1953.

[R] L. I. Rubenstein, *The Stefan Problem*, Transl. Math. Monographs 27, AMS, Providence, RI, 1971.

[RSP] D. S. Ryley, F. T. Smith, and G. Poots, *The inward solidification of spheres and circular cylinders*, Internat. J. Heat Mass Transfer, 17 (1974), pp. 1507–1516.

[S] A. M. Soward, *A unified approach to Stefan's problem for spheres and cylinders*, Proc. Royal Soc. London Ser. A, 373 (1980), pp. 131–147.

[SW] K. Stewartson and R. T. Waechter, *On Stefan's problem for spheres*, Proc. Royal Soc. London Ser. A, 348 (1976), pp. 415–526.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 33–48, January 1997

002

SELF-SIMILAR SOLUTIONS OF BARENBLATT'S MODEL FOR TURBULENCE*

JOSEPHUS HULSHOF[†]

Abstract. In this paper, we consider Barenblatt's k–ϵ model for turbulence. For the case of equal diffusion coefficients α and β, Barenblatt found explicit compactly supported self-similar solutions. From these, we obtain compactly supported solutions for $\alpha \neq \beta$ by transforming the equations into a four-dimensional quadratic system and verifying a transversality condition for a saddle-point connection. This involves the Poincaré transformation as well as classical properties of the hypergeometric equation and its solutions.

Key words. turbulence, compactly supported similarity solutions, quadratic systems, critical points at infinity, Poincaré transformation, saddle-point connections, transversality

AMS subject classification. 35K65

PII. S0036141095290033

Introduction. In this paper, we consider the system

$$\text{(KE)} \quad \begin{cases} k_t = \alpha\Big(\dfrac{k^2}{\varepsilon} k_x\Big)_x - \varepsilon, \\ \varepsilon_t = \beta\Big(\dfrac{k^2}{\varepsilon}\varepsilon_x\Big)_x - \gamma\dfrac{\varepsilon^2}{k}. \end{cases}$$

Here α, β, and γ are positive parameters and k and ε are unknown nonnegative functions of x (space) and t (time). This system is called the k–ε model and describes the evolution of turbulent bursts [B] (see also [LS], [HP], and [KV]); k stands for the turbulent energy density and ε is the dissipation rate of turbulent energy. In applications, α and β are usually different [LMRS, HL]. The model is also refered to in the literature as the b–ε model, which is, in fact, the original notation due to Kolmogorov ($k = b$) [K, P, MY]. We note that (KE) is a coupled system of two quasilinear diffusion-absorption equations. The diffusion coefficients may, depending on k and ε, become degenerate (very small) or singular (very large), and the second absorption term is also singular.

The only results that have been rigorously established so far are for the case where $\alpha = \beta$: for $\gamma > 3/2$, a family of explicit self-similar compactly supported "source-type" solutions was found by Barenblatt et al. [BGL], and for $\gamma > 1$, an existence result for solutions to the Cauchy problem was proved by Bertsch, Dal Passo, and Kersner [BdPK1, BdPK2], who also showed that for $\gamma > 3/2$, the self-similar solutions describe the intermediate asymptotics of these solutions.

This paper is concerned with the existence of compactly supported self-similar solutions when $\alpha \neq \beta$. Let us recall that the Barenblatt solutions are obtained by

*Received by the editors August 7, 1995; accepted for publication (in revised form) October 18, 1995. This research was supported by the Netherlands Organization for Scientific Research (NWO) and by EEC grant SC1-0019-C-(TT).

http://www.siam.org/journals/sima/28-1/29003.html.

[†]Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands (hulshof@wi.leidenuniv.nl).

substituting

$$k = \frac{A^2}{t^{2\mu}} f(\zeta), \qquad \varepsilon = \frac{A^2}{t^{2\mu+1}} g(\zeta), \qquad \zeta = \frac{x}{At^{1-\mu}}, \tag{0.1}$$

where $A > 0$ is a free-scaling parameter and where we restrict our attention to the case where $0 < \mu < 1$. Thus we look at profiles which decay and spread out as time evolves.

The equations for f and g are

$$\begin{cases} \alpha\Big(\dfrac{f^2}{g} f'\Big)' + (1-\mu)\zeta f' + 2\mu f - g = 0; & (0.2) \\ \beta\Big(\dfrac{f^2}{g} g'\Big)' + (1-\mu)\zeta g' + (1+2\mu)g - \gamma\dfrac{g^2}{f} = 0. & (0.3) \end{cases}$$

If we assume that

$$g(\zeta) = \kappa f(\zeta), \tag{0.4}$$

equations (0.2) and (0.3) can be reduced to one single equation if and only if

$$\alpha = \beta, \qquad \kappa = \frac{1}{\gamma - 1}, \tag{0.5}$$

the resulting equation for f being

$$\frac{\alpha}{\kappa}(ff')' + (1-\mu)\zeta f' + (2\mu - \kappa) f = 0. \tag{0.6}$$

Finally, if also

$$\mu = \frac{\kappa + 1}{3}, \quad 0 < \mu < 1, \tag{0.7}$$

equation (0.6) can be written as

$$\frac{3\alpha}{\kappa(2-\kappa)}(ff')' + (\zeta f)' = 0, \tag{0.8}$$

which has compactly supported nonnegative solutions

$$f(\zeta) = \left(C - \frac{\kappa(2-\kappa)}{6\alpha}\zeta^2\right)_+, \quad C > 0, \tag{0.9}$$

if and only if

$$0 < \kappa < 2. \tag{0.10}$$

Note that (0.5), (0.7), and (0.10) imply that $\gamma > \frac{3}{2}$.

We observe that (0.9) corresponds to the well-known Barenblatt profile for the porous-medium equation (denoted by (PME)) $u_t = (u^m)_{xx}$ with $m = 2$. In fact, substitution of $\varepsilon = \kappa k$ together with (0.5) reduces the full system (KE) to (PME).

Just as in the case of the (PME) (see, e.g., [A]), we see that at the boundary of the support of the solutions, the fluxes vanish, i.e.,

$$\frac{f^2}{g}f' \to 0 \quad \text{and} \quad \frac{f^2}{g}g' \to 0. \tag{0.11}$$

The main result of this paper is a perturbation of the explicit family of compactly supported similarity solutions above, yielding a similar family of solutions for $\gamma > 3/2$ and α close to β. This is an important and strong indication that the PDE results mentioned above for $\alpha = \beta$ are not isolated but really a first step towards a full theory for (KE).

THEOREM. *There exists an open neighborhood $\mathcal{O}$ of the set*

$$\left\{(\alpha,\beta,\gamma):\ \alpha=\beta>0,\ \gamma>\frac{3}{2}\right\}$$

such that for every $(\alpha,\beta,\gamma)\in\mathcal{O}$, there is precisely one $0<\mu<1$ for which equations (0.2) and (0.3) have a solution pair (f,g) with f and g symmetric and positive on $(-1,1)$ and

$$f(\zeta)\to 0,\quad g(\zeta)\to 0,\quad \frac{f(\zeta)}{g(\zeta)}f'(\zeta)\to-\alpha(1-\mu),\quad \frac{f(\zeta)^2}{g(\zeta)^2}g'(\zeta)\to-\beta(1-\mu) \tag{0.12}$$

as $\zeta\uparrow 1$. Moreover, if we write

$$\kappa=\frac{g(0)}{f(0)},\qquad \lambda=\frac{\alpha}{\beta}, \tag{0.13}$$

then in $\lambda=1$,

$$\kappa=\frac{1}{\gamma-1},\qquad \frac{d\mu}{d\lambda}=0,\qquad \frac{d\kappa}{d\lambda}=\frac{\kappa(2-\kappa)}{\kappa+1}\left(\kappa-1+\frac{2}{B_\kappa}\right). \tag{0.14}$$

Here B_κ is defined by

$$B_\kappa=\frac{\Gamma(\frac{1}{2})}{\Gamma(a)\Gamma(b)},\qquad a+b=\frac{3}{2},\qquad ab=\frac{3}{2(2-\kappa)}. \tag{0.15}$$

In order to perform the perturbation argument, we adapt the methods in [H] and introduce

$$t=\log\zeta,\quad x=\frac{\zeta f'}{f},\quad y=\frac{\zeta g'}{g},\quad z=\zeta^2\frac{g}{\alpha f^2},\quad u=\frac{g}{f}, \tag{0.16}$$

which transforms the two coupled nonautonomous second-order equations (0.2) and (0.3) into the four-dimensional first-order quadratic autonomous system

$$\text{(Q)}\quad\begin{cases}\dfrac{dx}{dt}=x(1-3x+y)-z(x(1-\mu)+2\mu-u);\\[2mm] \dfrac{dy}{dt}=y(1-2x)-\lambda z(y(1-\mu)+2\mu+1-\gamma u);\\[2mm] \dfrac{dz}{dt}=z(2+y-2x);\\[2mm] \dfrac{du}{dt}=u(y-x).\end{cases}$$

In section 1, we investigate this system. We find that symmetric profiles (f, g) correspond to the two-dimensional "fast" unstable manifold $\mathcal{F}$ of the positive u-axis and that the profiles satisfying the so-called interface condition as $0 < \zeta \uparrow \zeta^* < \infty$ are contained in the two-dimensional stable manifold $\mathcal{S}$ of a critical point at infinity on the line with direction vector

$$\begin{pmatrix} -(1-\mu) \\ -\lambda(1-\mu) \\ 1 \\ 0 \end{pmatrix}. \tag{0.17}$$

This involves the Poincaré transformation of (Q) and is carried out with the help of Maple. As a byproduct here, we find that it is necessary to assume that

$$\alpha \leq 2\beta \tag{0.18}$$

because otherwise $\mathcal{S}$ is one dimensional and contained in "infinity."

It follows from the analysis in section 1 that the compactly supported profiles we are looking for correspond to intersections of $\mathcal{F}$ and $\mathcal{S}$. In particular, and just as in [H], the explicit solutions above correspond to an orbit which is simply the straight line

$$x = y = -(1-\mu)z, \quad u = \kappa. \tag{0.19}$$

In section 2, we show that in the full $(x, y, z, u, \alpha, \beta, \gamma, \mu)$-space, the intersection of $\mathcal{F}$ and $\mathcal{S}$ is transversal at (0.19), thus obtaining our perturbation result. The dynamical-systems methods we use here were applied earlier in [AV] and [HV] to two-dimensional systems that come from scalar diffusion equations. However, in our case, the computations in which the hypergeometric function, the Gauss formula, and the Kummer relations appear [L] are much more involved, and again it is thanks to the help of Maple that we were able to pull through.

1. The quadratic system. In this section, we examine system (Q) in relation to the boundary conditions imposed on f and g. We note that every solution of (0.2)–(0.3) is mapped into an orbit of (Q) and that scaling with the parameter A in (0.1) corresponds to a shift in t.

By standard ODE theory [CL], there exists for every $p, q > 0$ a unique local solution (f, g) of (0.2)–(0.3) satisfying the initial conditions

$$f(0) = p, \quad g(0) = q, \quad f'(0) = 0, \quad g'(0) = 0. \tag{1.1}$$

This provides us with a two-parameter family of local solutions of (0.2)–(0.3). For the corresponding solution curve $S(t) = (x(t), y(t), z(t), u(t))$, we find

$$\lim_{t \downarrow -\infty} x(t)e^{-2t} = \lim_{\zeta \downarrow 0} \frac{f'(\zeta)}{\zeta f(\zeta)} = \frac{f''(0)}{f(0)} = \frac{1}{\alpha}(q - 2\mu p)\frac{q}{p^3}. \tag{1.2}$$

Here we have used (0.2) to compute $f''(0)$. Similarly, we find

$$\lim_{t \downarrow -\infty} y(t)e^{-2t} = \frac{1}{\beta}(\gamma q - (2\mu + 1)p)\frac{q}{p^3}, \qquad \lim_{t \downarrow -\infty} z(t)e^{-2t} = \lim_{\zeta \downarrow 0} \frac{g(\zeta)}{f(\zeta)^2} = \frac{\alpha q}{p^2}, \tag{1.3}$$

and, using l'Hôpital's rule,

$$\begin{aligned}
&\lim_{t\downarrow-\infty}\left(u(t)-\frac{q}{p}\right)e^{-2t}=\lim_{\zeta\downarrow 0}\frac{pg(\zeta)-qf(\zeta)}{p\zeta^2 f(\zeta)}=\lim_{\zeta\downarrow 0}\frac{pg'(\zeta)-qf'(\zeta)}{2p\zeta f(\zeta)+p\zeta^2 f'(\zeta)}\\
&=\lim_{\zeta\downarrow 0}\frac{pg''(\zeta)-qf''(\zeta)}{2pf(\zeta)+4p\zeta f'(\zeta)+p\zeta^2 f''(\zeta)}=\frac{q^2}{2p^4}\left(\frac{1}{\beta}(\gamma q-(2\mu+1)p)-\frac{1}{\alpha}(q-2\mu p)\right).
\end{aligned} \tag{1.4}$$

Thus $S(t)$ comes out of the point $(x,y,z,u)=(0,0,0,q/p)$ on the positive u-axis into the (invariant) open "quadrant" $O^+=\{z>0,u>0\}$ along an eigenvector corresponding to the eigenvalue 2 of the linearization of (Q) around $(0,0,0,q/p)$, which is

$$\begin{pmatrix}1 & 0 & -(2\mu-\kappa) & 0\\ 0 & 1 & -\lambda(2\mu+1-\gamma\kappa) & 0\\ 0 & 0 & 2 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}. \tag{1.5}$$

Here

$$\kappa=\frac{q}{p},\qquad \lambda=\frac{\alpha}{\beta}. \tag{1.6}$$

Clearly, the positive symmetric solution pairs (f,g) are mapped into the "fast unstable manifold" of the u-axis, the sheet of integral curves tangent to the eigenvector of 2. Note that the ratio κ determines the orbit.

Next, we consider solutions of (0.2)–(0.3) with $f(\zeta)\to 0$ and $g(\zeta)\to 0$ and satisfying the no-flux condition (0.11) as $\zeta\uparrow 1$. This cannot be viewed as an initial- (or final-) boundary value problem in such a straightforward manner as above, and therefore we turn to the quadratic system. Any such solution with both components decreasing to zero as $\zeta\uparrow 1$ is mapped into an orbit which escapes to infinity in finite time. Indeed, all of the other orbits contain solutions $S(t)=(x(t),y(t),z(t),u(t))$ which persist as $t\uparrow\infty$, and it is easy to see that the corresponding solutions (f,g) are positive in $\zeta=1$. Thus we look for orbits escaping to infinity in finite time with $x<0$, $y<0$, $z>0$, and $u>0$. This means that x and y cannot both be bounded.

For the study of the unbounded orbits, we use the Poincaré transformation to determine the critical points at infinity. Rewriting (Q) as

$$\text{(Q)}\quad\begin{cases}\dot{x}_1=P_1(x_1,x_2,x_3,x_4);\\ \dot{x}_2=P_2(x_1,x_2,x_3,x_4);\\ \dot{x}_3=P_3(x_1,x_2,x_3,x_4);\\ \dot{x}_4=P_4(x_1,x_2,x_3,x_4),\end{cases}$$

where $(x_1,x_2,x_3,x_4)=(x,y,z,u)$ and dots denote differentiation with respect to t, we introduce the new coordinates X_1,X_2,X_3,X_4, and V as follows:

$$x_i=\frac{X_i}{V}\quad(i=1,2,3,4),\qquad X_1^2+X_2^2+X_3^2+X_4^2+V^2=1. \tag{1.7}$$

This transforms (Q) into an autonomous polynomial system of five first-order differential equations for X_1,X_2,X_3,X_4, and V, which leaves the 4-sphere $S^4=\{X_1^2+X_2^2+X_3^2+X_4^2+V^2=1\}$ invariant.

Differentiating (1.7), we have

$$V\dot{V}+\sum_{j=1}^{4}X_j\dot{X}_j=0 \tag{1.8}$$

and

$$\dot{X}_iV-X_i\dot{V}=P_i^*, \tag{1.9}$$

where

$$P_i^*(X_1,X_2,X_3,X_4,V)=V^2P_i(x_1,x_2,x_3,x_4). \tag{1.10}$$

Thus the P_i^*'s are homogeneous polynomials of degree 2. Combining (1.8) and (1.9), we obtain

$$V\left(V^2+\sum_{j=1}^{4}X_j^2\right)\dot{V}=-V\sum_{j=1}^{4}X_jP_j^* \tag{1.11}$$

and, with (1.8) again,

$$\begin{aligned} V\left(V^2+\sum_{j=1}^{4}X_j^2\right)\dot{X}_i &= \left(V^2+\sum_{j=1}^{4}X_j^2\right)P_i^*+X_i\left(V^2+\sum_{j=1}^{4}X_j^2\right)\dot{V} \\ &= V^2P_i^*+\sum_{j=1}^{4}X_j(X_jP_i^*-X_iP_j^*). \end{aligned} \tag{1.12}$$

Thus integral curves of (Q) correspond to integral curves with $V>0$ on S^4 of the system

$$(\tilde{\mathrm{Q}})\quad\begin{cases} X_i'=V^2P_i^*+\sum_{j=1}^{4}X_j(X_jP_i^*-X_iP_j^*) \quad (i=1,2,3,4); \\ V'=-V\sum_{j=1}^{4}X_jP_j^*. \end{cases}$$

Here we have absorbed the factor

$$V\left(V^2+\sum_{j=1}^{4}X_j^2\right)$$

in the derivative.

Unbounded solutions of (Q) correspond to solutions of $(\tilde{\mathrm{Q}})$ which approach the invariant set $S^4\cap\{V=0\}$. The critical points "at infinity" of (Q) are by definition the critical points of $(\tilde{\mathrm{Q}})$ on $S^4\cap\{V=0\}$, which in turn are the solutions of

$$\begin{cases} \sum_{j=1}^{4}X_j(X_jP_i^*-X_iP_j^*)=0 \quad (i=1,2,3,4); & (1.13) \\ X_1^2+X_2^2+X_3^2+X_4^2+V^2=1. & (1.14) \end{cases}$$

Note that we have five equations for four unknowns. It is implicit in the Poincaré transformation that these equations are dependent.

Using Maple and again writing X, Y, Z, and U for X_1, X_2, X_3, and X_4, we find that (1.13) is equivalent to

$$\begin{cases} \bigl(Y^3 + (\lambda - 1)(1-\mu)Y^2Z - \gamma\lambda YZU - (1 - \mu Z^3 - (1-\mu)ZU^2\bigr)X \\ \qquad\qquad + ZU(Z^2 + Y^2 + U^2) + (Y^2 + 2U^2 + Z^2)X^2 = 0, \\ \bigl(XY - Y^2 - (1-\mu)(\lambda - 1)YZ + \gamma\lambda ZU\bigr)X^2 - (U^2 + ZU)XY \\ \qquad - (U^2 + Z^2)Y^2 - \lambda(1-\mu)(Z^2 + U^2)YZ + \gamma\lambda ZU(U^2 + Z^2) = 0, \\ Z\bigl(X^3 + (1-\mu)ZX^2 - (ZU + U^2)X + Y^3 + \lambda(1-\mu)Y^2Z - \gamma\lambda YZU\bigr) = 0, \\ U\bigl(2X^3 + (1-\mu)X^2Z + (Y^2 + Z^2 - ZU)X + Y^3 + \lambda(1-\mu)Y^2Z - \gamma\lambda YZU\bigr) = 0, \end{cases}$$

which at first sight looks too complicated to evaluate. However, if we multiply the third equation by Z and the fourth equation by U, subtraction gives

$$XZU(X^2 + Y^2 + Z^2 + U^2) = XZU = 0, \tag{1.15}$$

which reduces the system. Also, if we substitute $X = 0$ in the first equation, we obtain

$$ZU(Y^2 + Z^2 + U^2) = ZU = 0. \tag{1.16}$$

Thus all the solutions of (1.13)–(1.14) have either $Z = 0$ or $U = 0$ or have both. This allows us to solve (1.13)–(1.14) explicitly, either by hand or by again using Maple. The solutions (X, Y, Z, U) with $Z \geq 0$ and $U \geq 0$ are

$$(\pm 1, 0, 0, 0), \quad (0, \pm 1, 0, 0), \quad (0, 0, 1, 0), \quad (0, 0, 0, 1), \quad \left(\pm\sqrt{\frac{1}{2}}, \pm\sqrt{\frac{1}{2}}, 0, 0\right),$$

$$\left(\frac{-(1-\mu)}{\sqrt{1 + (1-\mu)^2}}, 0, \frac{1}{\sqrt{1 + (1-\mu)^2}}, 0\right), \quad \left(0, \frac{-\lambda(1-\mu)}{\sqrt{1 + \lambda^2(1-\mu)^2}}, \frac{1}{\sqrt{1 + \lambda^2(1-\mu)^2}}, 0\right),$$

and, last but not least,

$$P = \left(\frac{-(1-\mu)}{\sqrt{1 + (1+\lambda^2)(1-\mu)^2}}, \frac{-\lambda(1-\mu)}{\sqrt{1 + (1+\lambda^2)(1-\mu)^2}}, \frac{1}{\sqrt{1 + (1+\lambda^2)(1-\mu)^2}}, 0\right). \tag{1.17}$$

Solution curves of ($\tilde{\mathrm{Q}}$) going into P from $S^4 \cap \{V > 0\}$ correspond to solution curves of (Q) with

$$\frac{x}{z} \to -(1-\mu), \qquad \frac{y}{z} \to -\lambda(1-\mu), \qquad \frac{u}{z} \to 0 \tag{1.18}$$

so that in view of the equation for z,

$$\frac{d}{dt}\frac{1}{z(t)} \to (\lambda - 2)(1-\mu). \tag{1.19}$$

Thus if $\lambda < 2$, these orbits reach infinity in a finite time t^* with none of the functions $x(t)$, $y(t)$, and $z(t)$ integrable near t^*. (Note that $\lambda > 2$ is impossible, as the linearization of ($\tilde{\mathrm{Q}}$) around P will confirm.) Since

$$\int x(t)dt = \int \frac{f'(\zeta)}{f(\zeta)}d\zeta, \qquad \int y(t)dt = \int \frac{g'(\zeta)}{g(\zeta)}d\zeta, \tag{1.20}$$

it follows that

$$f(\zeta^*) = g(\zeta^*) = 0, \quad \zeta^* = e^{t^*}. \tag{1.21}$$

From (1.18), we also have
(1.22)

$$\frac{f(\zeta)}{g(\zeta)} f'(\zeta) = \frac{\alpha x(t)}{z(t)} e^t \to -\alpha(1-\mu)\zeta^*, \qquad \frac{f(\zeta)^2}{g(\zeta)^2} g'(\zeta) = \frac{\alpha y(t)}{z(t)} e^t \to -\beta(1-\mu)\zeta^*$$

so that condition (0.11) is satisfied at ζ^*. We shall call (1.22) the interface conditions for f and g.

The linearization of ($\tilde{\mathrm{Q}}$) around P has eigenvalues

$$-\frac{1-\mu}{\sqrt{1+(1+\lambda^2)(1-\mu)^2}}, \qquad -\frac{(1-\mu)(2-\lambda)}{\sqrt{1+(1+\lambda^2)(1-\mu)^2}}, \qquad 0,$$

$$\frac{1-\mu}{\sqrt{1+(1+\lambda^2)(1-\mu)^2}}, \qquad \frac{\lambda(1-\mu)}{\sqrt{1+(1+\lambda^2)(1-\mu)^2}},$$

i.e., up to a (positive if $\mu < 1$) multiple, simply

$$-1, \quad -(2-\lambda), \quad 0, \quad 1, \quad \lambda.$$

We note that zero is always an eigenvalue with eigenvector perpendicular to S^4. Since we only consider the flow on S^4, this eigenvector is irrelevant.

The only eigenvector with a nonzero V-component is the one corresponding to the eigenvalue which changes sign when λ crosses the value 2. Consequently, we may distinguish between two cases.

$0 < \lambda < 2$: The stable and unstable manifolds both have dimension two. The stable manifold contains a one-parameter family of solutions satisfying the interface conditions.

$\lambda > 2$: The stable manifold has dimension one and the unstable manifold dimension three. The stable manifold is contained in $\{V = 0\}$, implying that there are no orbits going into P coming from $\{V > 0\}$.

2. Transversality of the connection. In this section, we show that the explicit compactly supported solution which exists for

$$\alpha = \beta, \qquad \mu = \frac{\gamma}{3(\gamma-1)}, \qquad \gamma > \frac{3}{2}, \tag{2.1}$$

can be used to obtain a compactly supported solution for $\alpha \neq \beta$. Throughout this section, the value of $\gamma > 3/2$ is fixed. Condition (2.1) follows from (0.5) and (0.7).

The orbit of (Q) corresponding to the exact solutions in the introduction is the straight line (0.19), and it belongs to an analytic family of solution curves of the form

$$x = X(z;\kappa,\mu,\lambda), \qquad y = Y(z;\kappa,\mu,\lambda), \qquad u = U(z;\kappa,\mu,\lambda), \tag{2.2}$$

which are defined as the images under (0.16) of the symmetric solutions to (0.2)–(0.3), and together form the "fast unstable manifold" $\mathcal{F}$ of the u-axis. In particular, we have

$$X(0;\kappa,\mu,\lambda) = 0, \qquad Y(0;\kappa,\mu,\lambda) = 0, \qquad U(0;\kappa,\mu,\lambda) = \kappa = \frac{1}{\gamma-1}. \tag{2.3}$$

Thus we can use z and κ as a coordinate system on $\mathcal{F}$. Note that the analyticity of (2.2) excludes the other "slow" orbits coming out of the u-axis.

On the other hand, we have that at infinity the orbit (0.19) goes into the critical point P given by (1.17). Thus (0.19) also belongs to the stable manifold $\mathcal{S}$ of P. In the previous section, we have seen that $\mathcal{S}$ contains the similarity profiles satisfying the interface conditions and that its dimension is two. It can be written as a family of solutions of the form

$$x = X^*(z; c, \mu, \lambda), \qquad y = Y^*(z; c, \mu, \lambda), \qquad u = U^*(z; c, \mu, \lambda). \tag{2.4}$$

Here z and c are the parameters which can be used as a coordinate system on $\mathcal{S}$. We note that c is really given by the proof of the stable-manifold theorem and corresponds to a suitable smooth curve in the linearized stable manifold [Pe].

Both $\mathcal{F}$ and $\mathcal{S}$ are two dimensional. The straight line (0.19) lies in the intersection of $\mathcal{F}$ and $\mathcal{S}$. Since we are working in a four-dimensional space, the set of parameters for which this intersection is a curve should generically be a set of codimension one. To show that this is really the case in the vincinity of the exact solution above, we apply the implicit-function theorem to the following set of equations:

$$X(z; \kappa, \mu, \lambda) - X^*(z; c, \mu, \lambda) = 0; \tag{2.5}$$

$$Y(z; \kappa, \mu, \lambda) - Y^*(z; c, \mu, \lambda) = 0; \tag{2.6}$$

$$U(z; \kappa, \mu, \lambda) - U^*(z; c, \mu, \lambda) = 0. \tag{2.7}$$

Here the value of z can be taken fixed because the flow leaves $\mathcal{F}$ and $\mathcal{S}$ invariant.

In order to conclude that the solution set of (2.5)–(2.7) is of the form

$$\kappa = \kappa(\lambda), \qquad \mu = \mu(\lambda), \qquad c = c(\lambda), \tag{2.8}$$

we have to show that the matrix containing the partial derivatives of the left-hand sides with respect to κ, μ, and c has a nonzero determinant.

The functions $X(z)$, $Y(z)$, $U(z)$, $X^*(z)$, $Y^*(z)$, and $U^*(z)$ are solutions of the three-dimensional nonautonomous system obtained from (Q) by taking z as a new independent variable:

$$(\mathrm{Q}^*) \quad \begin{cases} \dfrac{dx}{dz} = \dfrac{x(1-3x+y) - z(x(1-\mu)+2\mu-u)}{z(2+y-2x)}; \\ \dfrac{dy}{dz} = \dfrac{y(1-2x) - \lambda z(y(1-\mu)+2\mu+1-\gamma u)}{z(2+y-2x)}; \\ \dfrac{du}{dz} = \dfrac{u(y-x)}{z(2+y-2x)}. \end{cases}$$

It follows from the proof of the stable-manifold theorem that we can compute the derivatives of these functions by differentiating (Q*) with respect to the parameters and solving the resulting equations under the appropiate boundary conditons.

Writing (Q*) as

$$\frac{d\xi}{dz} = H(\xi) = H(\xi; \mu, \lambda), \quad \xi(z) = (x(z), y(z), u(z)), \tag{2.9}$$

we have for the variation $d\xi(z) = (dx(z), dy(z), du(z))$ of ξ the equation

$$\frac{d}{dz} d\xi - \frac{\partial H}{\partial \xi} d\xi = dH = \frac{\partial H}{\partial \mu} d\mu + \frac{\partial H}{\partial \lambda} d\lambda. \tag{2.10}$$

In (2.10), the derivatives of H have to be evaluated at

$$x = y = -(1-\mu)z, \quad u = \kappa, \quad \mu = \frac{\kappa+1}{3}, \quad \gamma = \frac{\kappa+1}{\kappa}, \quad \lambda = 1. \tag{2.11}$$

Using Maple again, we find

$$\frac{\partial H}{\partial \xi} = \begin{pmatrix} \frac{1}{2z} + \frac{3}{2} \frac{2-\kappa}{6+(2-\kappa)z} & 0 & \frac{3}{6+(2-\kappa)z} \\ 0 & \frac{1}{2z} + \frac{3}{2} \frac{2-\kappa}{6+(2-\kappa)z} & \frac{3(\kappa+1)}{\kappa(6+(2-\kappa)z)} \\ \frac{-\kappa}{2z} + \frac{2-\kappa}{2} \frac{\kappa}{6+(2-\kappa)z} & \frac{\kappa}{2z} - \frac{2-\kappa}{2} \frac{\kappa}{6+(2-\kappa)z} & 0 \end{pmatrix}, \tag{2.12}$$

while

$$\frac{\partial H}{\partial \mu} = \begin{pmatrix} -1 \\ -1 \\ 0 \end{pmatrix}, \qquad \frac{\partial H}{\partial \lambda} = \begin{pmatrix} 0 \\ \frac{2-\kappa}{3} - \frac{3(2-\kappa)}{6+(2-\kappa)z} \\ 0 \end{pmatrix}. \tag{2.13}$$

In what follows, we shall compute the general solution of (2.9)–(2.13) explicitly in terms of hypergeometric functions. To do so we transform (2.9)–(2.13) by

$$w = \frac{(2-\kappa)z}{6+(2-\kappa)z}, \qquad z = \frac{6w}{(2-\kappa)(1-w)},$$

$$G(w) = dx(z), \qquad J(w) = dx(z) - dy(z), \qquad F(w) = du(z) \tag{2.14}$$

into

$$\begin{pmatrix} G'(w) \\ J'(w) \\ F'(w) \end{pmatrix} = \begin{pmatrix} \frac{1}{2w} + \frac{2}{1-w} & 0 & \frac{3}{2-\kappa} \frac{1}{1-w} \\ 0 & \frac{1}{2w} + \frac{2}{1-w} & -\frac{1}{\kappa} \frac{3}{2-\kappa} \frac{1}{1-w} \\ 0 & -\frac{\kappa}{2w} & 0 \end{pmatrix} \begin{pmatrix} G(w) \\ J(w) \\ F(w) \end{pmatrix}$$

$$- \frac{6}{2-\kappa} \frac{1}{(1-w)^2} \begin{pmatrix} d\mu \\ 0 \\ 0 \end{pmatrix} + \frac{1-3w}{(1-w)^2} \begin{pmatrix} 0 \\ d\lambda \\ 0 \end{pmatrix}. \tag{2.15}$$

For $F(w)$, this yields

$$w(1-w) \frac{d^2}{dw^2} F(w) + \left(\frac{1}{2} - \frac{5}{2} w \right) \frac{d}{dw} F(w) - \frac{3}{2(2-\kappa)} F(w) = \frac{\kappa(3w-1)}{2(1-w)} d\lambda, \tag{2.16}$$

which has an explicit particular solution, namely,

$$\kappa(2-\kappa) \left(1 - \frac{2}{\kappa+1} \frac{1}{1-w} \right) d\lambda. \tag{2.17}$$

The homogeneous part of (2.16) is the standard hypergeometric equation

$$w(1-w) f''(w) + (c - (1+a+b)w) f'(w) - ab f(w) = 0 \tag{2.18}$$

with parameters a, b, and c given by

$$a+b=\frac{3}{2}, \qquad ab=\frac{3}{2(2-\kappa)}, \qquad \text{and} \quad c=\frac{1}{2}. \tag{2.19}$$

The general solution of (2.18) is given by

$$f(w)=C_1F_1(w)+C_2F_2(w), \tag{2.20}$$

where

$$F_1(w)=F\left(a,b;\frac{1}{2};w\right)=\frac{1}{1-w}F\left(\frac{1}{2}-a,\frac{1}{2}-b;\frac{1}{2};w\right) \tag{2.21}$$

and

$$F_2(w)=w^{\frac{1}{2}}F\left(\frac{1}{2}+a,\frac{1}{2}+b;\frac{3}{2};w\right)=\frac{w^{\frac{1}{2}}}{1-w}F\left(1-a,1-b;\frac{3}{2};w\right). \tag{2.22}$$

Consequently, the general solution of the homogeneous part of (2.15) is given by

$$\begin{pmatrix} G_{\text{hom}}(w) \\ J_{\text{hom}}(w) \\ F_{\text{hom}}(w) \end{pmatrix} = C_1\begin{pmatrix} 2wF_1'(w) \\ -\frac{2w}{\kappa}F_1'(w) \\ F_1(w) \end{pmatrix} + C_2\begin{pmatrix} 2wF_2'(w) \\ -\frac{2w}{\kappa}F_2'(w) \\ F_2(w) \end{pmatrix} + C_3\begin{pmatrix} \frac{w^{\frac{1}{2}}}{(1-w)^2} \\ 0 \\ 0 \end{pmatrix}. \tag{2.23}$$

The hypergeometric part in (2.23) can be derived from the special form of the matrix in (2.15).

A particular solution of (2.15) is

$$\begin{pmatrix} G_p(w) \\ J_p(w) \\ F_p(w) \end{pmatrix} = \begin{pmatrix} \frac{2\kappa w}{(1-w)^2}(3\frac{\kappa-1}{\kappa+1}-w) \\ \frac{4(2-\kappa)}{\kappa+1}\frac{w}{(1-w)^2} \\ \kappa(2-\kappa)(1-\frac{2}{\kappa+1}\frac{1}{1-w}) \end{pmatrix} d\lambda + \begin{pmatrix} \frac{-12}{2-\kappa}\frac{w}{(1-w)^2} \\ 0 \\ 0 \end{pmatrix} d\mu. \tag{2.24}$$

Thus the general solution of (2.15) is the sum of (2.23) and (2.24).

We can now write the partial derivatives of (2.2) for (2.11). The analyticity near $z=0$ combined with (2.3) implies that we have to take

$$C_2=C_3=0, \quad C_1+\kappa(2-\kappa)\frac{\kappa-1}{\kappa+1}d\lambda=d\kappa \tag{2.25}$$

so that

$$\begin{pmatrix} \frac{\partial X}{\partial \kappa} & \frac{\partial X}{\partial \mu} & \frac{\partial X}{\partial \lambda} \\ \frac{\partial (X-Y)}{\partial \kappa} & \frac{\partial (X-Y)}{\partial \mu} & \frac{\partial (X-Y)}{\partial \lambda} \\ \frac{\partial U}{\partial \kappa} & \frac{\partial U}{\partial \mu} & \frac{\partial U}{\partial \lambda} \end{pmatrix} = \tag{2.26}$$

$$\begin{pmatrix} 2wF'(a,b;\frac{1}{2};w) & \frac{-12}{2-\kappa}\frac{w}{(1-w)^2} & \frac{2\kappa w}{(1-w)^2}(3\frac{\kappa-1}{\kappa+1}-w)+2\kappa(2-\kappa)\frac{1-\kappa}{\kappa+1}wF'(a,b;\frac{1}{2};w) \\ -\frac{2w}{\kappa}F'(a,b;\frac{1}{2};w) & 0 & \frac{4(2-\kappa)}{\kappa+1}\frac{w}{(1-w)^2}-2(2-\kappa)\frac{1-\kappa}{\kappa+1}wF'(a,b;\frac{1}{2};w) \\ F(a,b;\frac{1}{2};w) & 0 & \kappa(2-\kappa)(1-\frac{2}{\kappa+1}\frac{1}{1-w})+\kappa(2-\kappa)\frac{1-\kappa}{\kappa+1}F(a,b;\frac{1}{2};w) \end{pmatrix}.$$

Next, we compute the partial derivatives of (2.4). The boundary conditions are now at $z = \infty$ and follow from (1.17), which implies that

$$\frac{dx(z)}{z} \to d\mu, \quad \frac{dy(z)}{z} \to d\mu - \frac{2-\kappa}{3} d\lambda, \quad \frac{du(z)}{z} \to 0 \quad \text{as } z \to \infty, \tag{2.27}$$

equivalent (recall (2.14)) to

$$\lim_{w\uparrow 1}(1-w)G(w) = \frac{6}{2-\kappa} d\mu, \qquad \lim_{w\uparrow 1}(1-w)J(w) = 2d\lambda, \qquad \lim_{w\uparrow 1}(1-w)F(w) = 0. \tag{2.28}$$

In order to choose the constants C_1, C_2, and C_3 accordingly, we need the asymptotic expansions of (2.23)–(2.24) as $w \uparrow 1$. At first glance, the reader may want to skip these calculations and proceed directly to (2.47).

We note that Gauss's formula implies that

$$\lim_{w\uparrow 1}(1-w)F_1(w) = B_\kappa = \frac{\Gamma(\frac{1}{2})}{\Gamma(a)\Gamma(b)}, \qquad \lim_{w\uparrow 1}(1-w)F_2(w) = A_\kappa = \frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{1}{2}+a)\Gamma(\frac{1}{2}+b)}, \tag{2.29}$$

i.e.,

$$F_1(w) = \frac{B_\kappa}{1-w} + o\left(\frac{1}{1-w}\right), \qquad F_2(w) = \frac{A_\kappa}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w \uparrow 1. \tag{2.30}$$

For the corresponding first components of the homogeneous solution, we find

$$\begin{aligned} 2wF_1'(w) &= 2w\frac{ab}{\frac{1}{2}}F\left(a+1, b+1; \frac{3}{2}; w\right) = \frac{6}{2-\kappa}\frac{w}{(1-w)^2}F\left(\frac{1}{2}-a, \frac{1}{2}-b; \frac{3}{2}; w\right) \\ &= \frac{6}{2-\kappa}\frac{w}{(1-w)^2} \\ &\quad \times \left(\frac{\Gamma(\frac{3}{2})\Gamma(2)}{\Gamma(a+1)\Gamma(b+1)} - \frac{(\frac{1}{2}-a)(\frac{1}{2}-b)}{\frac{3}{2}}\frac{\Gamma(\frac{5}{2})\Gamma(1)}{\Gamma(a+1)\Gamma(b+1)}(1-w) \right. \\ &\qquad\qquad \left. + o(1-w)\right) \\ &= 2B_\kappa\left(\frac{1}{(1-w)^2} - \frac{w}{(1-w)^2}\right)\left(1 + \left(\frac{1}{2} - \frac{3}{2(2-\kappa)}\right)(1-w) + o(1-w)\right) \\ &= B_\kappa\left(\frac{2}{(1-w)^2} - \left(1 + \frac{3}{2-\kappa}\right)\frac{1}{1-w} + o(\frac{1}{1-w})\right) \quad \text{as } w \uparrow 1. \end{aligned} \tag{2.31}$$

In this computation, we have used the Gauss relation for both $F(1/2 - a, 1/2 - b; 3/2; w)$ and its derivative. Similarly, we have

$$\begin{aligned} 2wF_2'(w) &= w^{\frac{1}{2}}F\left(\frac{1}{2}+a, \frac{1}{2}+b; \frac{3}{2}; w\right) + 2w^{\frac{3}{2}}\frac{(\frac{1}{2}+a)(\frac{1}{2}+b)}{\frac{3}{2}}F\left(\frac{3}{2}+a, \frac{3}{2}+b; \frac{5}{2}; w\right) \\ &= F_2(w) + 2w^{\frac{3}{2}}\frac{(\frac{1}{2}+a)(\frac{1}{2}+b)}{\frac{3}{2}}\frac{1}{(1-w)^2}F\left(1-a, 1-b; \frac{5}{2}; w\right) \\ &= F_2(w) + \frac{2w^{\frac{3}{2}}}{(1-w)^2} \end{aligned} \tag{2.32}$$

$$\begin{aligned}
&\times \frac{(\frac{1}{2}+a)(\frac{1}{2}+b)}{\frac{3}{2}}\Bigg(\frac{\Gamma(\frac{5}{2})\Gamma(2)}{\Gamma(\frac{3}{2}+a)\Gamma(\frac{3}{2}+b)} - \frac{(1-a)(1-b)}{\frac{5}{2}}\frac{\Gamma(\frac{7}{2})\Gamma(1)}{\Gamma(\frac{3}{2}+a)\Gamma(\frac{3}{2}+b)}(1-w) \\
&\qquad + o(1-w)\Bigg) \\
&\quad = F_2(w) + \frac{2w^{\frac{3}{2}}}{(1-w)^2}A_\kappa\left(1+\left(\frac{1}{2}-\frac{3}{2(2-\kappa)}\right)(1-w)+o(1-w)\right) \\
&\quad = \frac{A_\kappa}{1-w} + o\left(\frac{1}{1-w}\right) \\
&+ \frac{2}{(1-w)^2}\left(1-\frac{3}{2}(1-w)+o(1-w)\right)A_\kappa\Bigg(1+\left(\frac{1}{2}-\frac{3}{2(2-\kappa)}\right)(1-w) \\
&\qquad + o(1-w)\Bigg) \\
&\quad = A_\kappa\left(\frac{2}{(1-w)^2}-\left(1+\frac{3}{2-\kappa}\right)\frac{1}{1-w}+o\left(\frac{1}{1-w}\right)\right) \quad \text{as } w\uparrow 1.
\end{aligned}$$

For the first component corresponding to C_3, we have

$$\text{(2.33)} \qquad \frac{w^{\frac{1}{2}}}{(1-w)^2} = \frac{1}{(1-w)^2} - \frac{1}{2}\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1.$$

For the particular solution corresponding to $d\lambda$, the third component is

$$\text{(2.34)} \quad \kappa(2-\kappa)\left(1-\frac{2}{\kappa+1}\frac{1}{1-w}\right) = -\frac{2\kappa(2-\kappa)}{\kappa+1}\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1,$$

the second component is

(2.35)

$$\frac{4(2-\kappa)}{\kappa+1}\frac{w}{(1-w)^2} = \frac{4(2-\kappa)}{\kappa+1}\frac{1}{(1-w)^2} - \frac{4(2-\kappa)}{\kappa+1}\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1,$$

and the first component is

(2.36)

$$\frac{2\kappa w}{(1-w)^2}\left(3\frac{\kappa-1}{\kappa+1}-w\right) = \frac{4\kappa(\kappa-2)}{\kappa+1}\frac{1}{(1-w)^2} + \frac{2\kappa(5-\kappa)}{\kappa+1}\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1.$$

Finally, for the first component of the particular solution for $d\mu$,

$$\text{(2.37)} \quad -\frac{12}{2-\kappa}\frac{w}{(1-w)^2} = -\frac{12}{2-\kappa}\frac{1}{(1-w)^2} + \frac{12}{2-\kappa}\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1.$$

Now that we have all of the asymptotic expansions as $w\uparrow 1$, we have to choose the constants in such a way that (2.28) is satisfied. First, we look at the third component,

$$\text{(2.38)} \quad \begin{aligned}
F(w) &= C_1F_1(w) + C_2F_2(w) + \kappa(2-\kappa)\left(1-\frac{2}{\kappa+1}\frac{1}{1-w}\right)d\lambda \\
&= \left(C_1B_\kappa + C_2A_\kappa - \frac{2\kappa(2-\kappa)}{\kappa+1}d\lambda\right)\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w\uparrow 1,
\end{aligned}$$

which forces us to take

$$C_1B_\kappa + C_2A_\kappa - \frac{2\kappa(2-\kappa)}{\kappa+1}d\lambda = 0. \tag{2.39}$$

Then by the Kummer relation,

$$2A_\kappa F_1(w) - 2B_\kappa F_2(w) = F(a,b;2;w), \tag{2.40}$$

$$F(w) = \kappa(2-\kappa)\left(\frac{1}{B_\kappa}\frac{2}{\kappa+1}F\left(a,b;\frac{1}{2};w\right) - \frac{2}{\kappa+1}\frac{1}{1-w} + 1\right)d\lambda + CF(a,b;2;w), \tag{2.41}$$

where $C = -C_2/(2B_\kappa)$.

For the second component, we then obviously have that the terms with $(1-w)^{-2}$ disappear and that

$$\begin{aligned} J(w) &= -\frac{2w}{\kappa}F'(w) = -C_1\frac{2w}{\kappa}F_1'(w) - C_2\frac{2w}{\kappa}F_2'(w)\frac{4(2-\kappa)}{\kappa+1}\frac{w}{(1-w)^2}d\lambda \\ &= \left(\frac{1}{\kappa}(C_1B_\kappa + C_2A_\kappa)\left(1+\frac{3}{2-\kappa}\right) - \frac{4(2-\kappa)}{\kappa+1}d\lambda\right)\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \\ &= \frac{2}{1-w}d\lambda + o\left(\frac{1}{1-w}\right) \quad \text{as } w \uparrow 1, \end{aligned} \tag{2.42}$$

which agrees with (2.28) and therefore gives no further restriction on the constants C_1, C_2, and C_3. Thus

$$J(w) = \frac{4(2-\kappa)w}{\kappa+1}\left(-\frac{1}{B_\kappa}F'\left(a,b;\frac{1}{2};w\right) + \frac{1}{(1-w)^2}\right)d\lambda - \frac{2Cw}{\kappa}F'(a,b;2;w), \tag{2.43}$$

Finally, for the first component, using (2.39) again,

$$\begin{aligned} G(w) &= 2wC_1F_1'(w) + 2wC_2F_2'(w) + C_3\frac{w^{\frac{1}{2}}}{(1-w)^2} \\ &\qquad + \frac{2\kappa w}{(1-w)^2}\left(3\frac{\kappa-1}{\kappa+1} - w\right)d\lambda - \frac{12}{2-\kappa}\frac{w}{(1-w)^2}d\mu \\ &= \left(C_3 - \frac{12}{2-\kappa}d\mu\right)\frac{1}{(1-w)^2} + \left(\frac{12}{2-\kappa}d\mu - \frac{1}{2}C_3\right)\frac{1}{1-w} + o\left(\frac{1}{1-w}\right) \quad \text{as } w \uparrow 1 \end{aligned} \tag{2.44}$$

so that

$$C_3 = \frac{12}{2-\kappa}d\mu \tag{2.45}$$

ensures that (2.28) holds. Thus

$$\begin{aligned} G(w) = &\left(\frac{4\kappa(2-\kappa)w}{(\kappa+1)B_\kappa}F'\left(a,b;\frac{1}{2};w\right) + \frac{2\kappa w}{(1-w)^2}\left(3\frac{\kappa-1}{\kappa+1} - w\right)\right)d\lambda \\ &+ 2CwF'(a,b;2;w) + \frac{12}{2-\kappa}\frac{w^{\frac{1}{2}}-w}{(1-w)^2}d\mu. \end{aligned} \tag{2.46}$$

From (2.41), (2.43), and (2.46), we then have for an appropiate choice of the coordinate c in (2.4) that

$$\begin{pmatrix} \frac{\partial X^*}{\partial c} & \frac{\partial X^*}{\partial \mu} & \frac{\partial X^*}{\partial \lambda} \\ \frac{\partial (X^*-Y^*)}{\partial c} & \frac{\partial (X^*-Y^*)}{\partial \mu} & \frac{\partial (X^*-Y^*)}{\partial \lambda} \\ \frac{\partial U^*}{\partial c} & \frac{\partial U^*}{\partial \mu} & \frac{\partial U^*}{\partial \lambda} \end{pmatrix} \tag{2.47}$$

$$= \begin{pmatrix} 2wF'(a,b;2;w) & \frac{12}{2-\kappa}\frac{w^{\frac{1}{2}}-w}{(1-w)^2} & \frac{2\kappa w}{(1-w)^2}(3\frac{\kappa-1}{\kappa+1}-w) + \frac{4\kappa(2-\kappa)w}{(\kappa+1)B_\kappa}F'(a,b;\frac{1}{2};w) \\ -\frac{2w}{\kappa}F'(a,b;2;w) & 0 & \frac{4(2-\kappa)}{\kappa+1}\frac{w}{(1-w)^2} - \frac{4(2-\kappa)w}{(\kappa+1)B_\kappa}F'(a,b;\frac{1}{2};w) \\ F(a,b;2;w) & 0 & \kappa(2-\kappa)(1-\frac{2}{\kappa+1}\frac{1}{1-w}) + \frac{2\kappa(2-\kappa)}{(\kappa+1)B_\kappa}F(a,b;\frac{1}{2};w) \end{pmatrix}.$$

Writing (2.5)–(2.7) as

$$\mathcal{F}(z;c,\kappa,\mu,\lambda) = \begin{pmatrix} X - X^* \\ X - X^* - Y + Y^* \\ U - U^* \end{pmatrix} = 0,$$

it follows that $\partial\mathcal{F}/\partial(c,\kappa,\mu,\lambda) =$

$$\begin{pmatrix} -2wF'(a,b;2;w) & 2wF'(a,b;\frac{1}{2};w) & -\frac{12}{2-\kappa}\frac{w^{\frac{1}{2}}}{(1-w)^2} & \frac{2\kappa(2-\kappa)}{\kappa+1}\beta_\kappa wF'(a,b;\frac{1}{2};w) \\ \frac{2w}{\kappa}F'(a,b;2;w) & -\frac{2w}{\kappa}F'(a,b;\frac{1}{2};w) & 0 & -\frac{2(2-\kappa)}{\kappa+1}\beta_\kappa wF'(a,b;\frac{1}{2};w) \\ -F(a,b;2;w) & F(a,b;\frac{1}{2};w) & 0 & \frac{\kappa(2-\kappa)}{\kappa+1}\beta_\kappa F(a,b;\frac{1}{2};w) \end{pmatrix}, \tag{2.48}$$

where

$$\beta_\kappa = 1 - \kappa - \frac{2}{B_\kappa}. \tag{2.49}$$

Clearly, the first three columns in this matrix have maximal rank for any $0 < w < 1$ because the Wronskian of the two hypergeometric functions $F(a,b;1/2;w)$ and $F(a,b;2;w)$ is nonzero. It follows that we can write the solution set of (2.5-7) in the form (2.8) with

$$\begin{pmatrix} -2wF'(a,b;2;w) & 2wF'(a,b;\frac{1}{2};w) & -\frac{12}{2-\kappa}\frac{w^{\frac{1}{2}}}{(1-w)^2} \\ \frac{2w}{\kappa}F'(a,b;2;w) & -\frac{2w}{\kappa}F'(a,b;\frac{1}{2};w) & 0 \\ -F(a,b;2;w) & F(a,b;\frac{1}{2};w) & 0 \end{pmatrix} \begin{pmatrix} \frac{dc}{d\lambda} \\ \frac{d\kappa}{d\lambda} \\ \frac{d\mu}{d\lambda} \end{pmatrix}$$

$$= -\begin{pmatrix} \frac{2\kappa(2-\kappa)}{\kappa+1}\beta_\kappa wF'(a,b;\frac{1}{2};w) \\ -\frac{2(2-\kappa)}{\kappa+1}\beta_\kappa wF'(a,b;\frac{1}{2};w) \\ \frac{\kappa(2-\kappa)}{\kappa+1}\beta_\kappa F(a,b;\frac{1}{2};w) \end{pmatrix}, \tag{2.50}$$

whence, using Cramer's rule,

$$\frac{dc}{d\lambda} = \frac{d\mu}{d\lambda} = 0, \qquad \frac{d\kappa}{d\lambda} = \frac{\kappa(2-\kappa)}{\kappa+1}\left(\kappa - 1 + \frac{2}{B_\kappa}\right). \tag{2.51}$$

Acknowledgments. I would like to thank G. I. Barenblatt, M. Bertsch, and L. A. Peletier for their encouragement, S. B. Angenent, G. Marí Beffa, F. Beukers, and

E. M. Opdam for fruitful discussions about special functions, the makers of Maple for making it all possible, and finally D. G. Aronson and J. L. Vazquez for [AV].

REFERENCES

[A] D. G. Aronson, *The porous medium equation*, in Some Problems in Nonlinear Diffusion, A. Fasano and M. Primicerio, eds., Lecture Notes in Math. 1224, Springer-Verlag, Berlin, 1986, pp. 1–46.

[AV] D. G. Aronson and J. L. Vazquez, *Calculation of anamolous exponents in nonlinear diffusion*, Phys. Rev. Lett., 72 (1994), pp. 348–351.

[B] G. I. Barenblatt, *Self-similar turbulence propagation from an instantaneous plane source*, in Nonlinear Dynamics and Turbulence, G. I. Barenblatt, G. Iooss, and D. D. Joseph, eds., Pitman, Boston, 1983, pp. 48–60.

[BGL] G. I. Barenblatt, N. L. Galerkina, and M. V. Luneva, *Evolution of a turbulent burst*, Inzh.-Fiz. Zh., 53 (1987), pp. 773–740 (in Russian).

[BdPK1] M. Bertsch, R. Dal Passo, and R. Kersner, *Parameter dependence in the $b-\varepsilon$ model*, Differential Integral Equations, 7 (1994), pp. 1195–1214.

[BdPK2] M. Bertsch, R. Dal Passo, and R. Kersner, *The evolution of turbulent bursts: The $b-\varepsilon$ model*, European J. Appl. Math., 5 (1994), pp. 537–557.

[CL] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, Krieger Publishing Company, Melbourne, FL, 1984.

[HL] K. Hanjalic and B. E. Launder, *A Reynolds stress model of turbulence and its applications to thin shear flows*, J. Fluid. Mech., 52 (1974), pp. 609–638.

[HP] S. P. Hastings and L. A. Peletier, *On the decay of turbulent bursts*, European J. Appl. Math., 3 (1992), pp. 319–341.

[H] J. Hulshof, *Similarity solutions of the porous medium equation with sign changes*, J. Math. Anal. Appl., 157 (1991), pp. 75–111.

[HV] J. Hulshof and J. L. Vazquez, *Selfsimilar solutions of the second kind for the modified porous medium equation*, European J. Appl. Math., 5 (1994), pp. 391–403.

[KV] S. Kamin and J. L. Vazquez, *The propagation of turbulent bursts*, European J. Appl. Math., 3 (1992), pp. 263–272.

[K] A. N. Kolmogorov, *Equation of turbulent motion of incompressible fluids*, Izv. Akad. Nauk SSSR, 6 (1942), pp. 56–58.

[L] N. N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.

[LMRS] B. E. Launder, A. P. Morse, W. Rodi, and D. B. Spalding, *Prediction of free shear flows: A comparison of six turbulence models*, NASA SP, 321 (1972).

[LS] B. E. Launder and D. B. Spalding, *The numerical computation of turbulent flows*, Comput. Math. Appl. Mech. Engrg., 3 (1974), pp. 269–289.

[MY] A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics*, Vols. 1 and 2, MIT Press, Cambridge, MA, 1971 and 1975.

[Pe] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, Berlin, 1991.

[P] L. Prandtl, *Über ein neues Formelsystem für die ausgebildete Turbulenz*, Nachr. Akad. Wiss. Göttingen Math.-Phys., K1 (1945), pp. 6–18.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 49–59, January 1997

ON UNIQUENESS OF RECOVERY OF THE DISCONTINUOUS CONDUCTIVITY COEFFICIENT OF A PARABOLIC EQUATION[*]

ALAEDDIN ELAYYAN[†] AND VICTOR ISAKOV[‡]

Abstract. We prove uniqueness of a discontinuous principal coefficient of a second-order parabolic equation of the form $a_0 + \chi(Q^*)b$ with known smooth a_0 and unknown $b = b(x)$ from all possible lateral boundary measurements of solutions of this equation. In the proofs, we make use of singular solutions of parabolic equations.

Key words. partial differential equations, inverse problems

AMS subject classification. 35R30

PII. S0036141095286010

Introduction. We consider the problem of recovery of the coefficient a of the parabolic equation

$$u_t - \operatorname{div}(a\nabla u) = 0 \quad \text{in } Q = \Omega \times (0, T)$$

with the initial and boundary conditions

$$u = 0 \quad \text{on } \Omega \times \{0\}, \qquad u = g \quad \text{on } \partial\Omega \times [0, T]$$

when $\partial u/\partial\nu$ is given for all (regular) g. Here Ω is a bounded domain in $\mathbb{R}^n$, $2 \le n$, with the boundary $\partial\Omega \in C^2$. In this paper, we prove uniqueness of discontinuous $a = a_0 + \chi(Q^*)b$, where $\chi(Q^*)$ is the indicator function of an open set $Q^* \subset Q$ with the Lipschitz lateral boundary $\partial_x Q^*$ changing with time and $a_0 = a_0(x)$ and $b = b(x)$ are, respectively, given and unknown $C^2(\bar{\Omega})$-functions. For elliptic equations, uniqueness was proven by Kohn and Vogelius [8] (piecewise-analytic a) and Isakov [5] (Lipschitz Q^* and smooth b). Also for elliptic equations, when one is making use of only one set of u, $\partial u/\partial\nu$ on $\partial\Omega$, some partial global uniqueness results for Q^* were obtained by Friedman and Isakov [4]. Regarding parabolic equations, we can refer only to Bellout's study [2] of local stability in the inverse problem. This inverse parabolic problem is fundamental for groundwater search [12] in particular and important for many engineering applications.

We introduce some notation. For standard notation, we refer to Friedman [3] and Ladyzhenskaja, Solonnikov, and Ural'ceva [9].

For an open set Q in the layer $\mathbb{R}^n \times (0, T)$, the lateral boundary $\partial_x Q$ is the x-boundary that is the closure of the set $\partial Q|\{t = 0 \text{ or } t = T\}$. We say that Q is x-Lipschitz if its x-boundary is locally the graph of a function $x_j = \gamma(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n, t)$ that is Lipschitz.

[*] Received by the editors May 12, 1995; accepted for publication October 23, 1995.
http://www.siam.org/journals/sima/28-1/28601.html

[†] Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033. Current address: Department of Mathematics and Computer Science, Birzeit University, P.O. Box 14, Birzeit, West Bank, Israel (elayyan@math.birzeit.edu).

[‡] Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033 (isakov@twsuvm.uc.twsu.edu). The research of this author was partly supported by National Science Foundation grants DMS-9101421 and DMS-9501510.

1. Statement of results. Let Γ_0 be $\partial\Omega \cup B_0$ for some ball B_0 centered at a point of $\partial\Omega$.

We are interested in finding an open set Q_j and a function b_j entering the parabolic initial-boundary value problem

$$(u_j)_t - \operatorname{div}(a_j \nabla u_j) = 0 \quad \text{in } Q, \tag{1.1}$$

$$u_j = g \quad \text{on } S = \partial\Omega \times (0,T), \tag{1.2}$$

$$u = 0 \quad \text{on } \Omega \times \{0\}, \tag{1.3}$$

where

$$a_j = a_0 + \chi(Q_j) b_j > \epsilon > 0, \quad b_j \neq 0 \text{ on } \partial Q_j. \tag{1.4}$$

It is well known that for any $g \in C^{2,1}(\bar{S})$, $g = g_t = g_{tt} = 0$ on $\partial\Omega \times \{0\}$, there is a unique (generalized) solution u_j of this problem and $u_j \in C^{\lambda}(\bar{Q})$ for some $\lambda \in (0,1)$, $\nabla_x u_j \in L_2(Q)$, and $\in C(\bar{Q}\backslash\bar{Q}_j)$. For this and for other results about the direct parabolic problem (1.1)–(1.4), we refer to Friedman [3] and Ladyzhenskaja, Solonnikov, and Ural'ceva [9, pp. 153, 204, and 227].

Our main result is the following theorem.

THEOREM 1.1. *Suppose Q_1 and Q_2 are open x-Lipschitz sets, $Q_j \subset \Omega \times (-T, 2T)$, and*

$$\text{the sets } (Q\backslash\overline{Q}_j) \cap \{t = \tau\} \text{ are connected} \quad \text{when } 0 < \tau < T. \tag{1.5}$$

If solutions u_j to the initial-boundary value problems (1.1), (1.2), *and* (1.3) *satisfy the equality*

$$\partial u_1/\partial\nu = \partial u_2/\partial\nu \quad \text{on } \Gamma_0 \times (0,T) \quad (\nu \text{ is a normal}) \tag{1.6}$$

for all $g \in C^2(\partial\Omega \times [0,T])$ with $\operatorname{supp} g \subset \Gamma_0 \times (0,T)$, then

$$a_1 = a_2 \quad \text{on } Q. \tag{1.7}$$

This result guarantees uniqueness of reconstruction of Q_j from all possible lateral measurements for an arbitrary $T > 0$.

The paper is organized as follows. In section 2, we will show that if equality (1.6) is valid for all Dirichlet boundary data, g implies certain integral relations which can be interpreted as orthogonality relations. To prove uniqueness in section 4, we will modify an approach from [5] (the use of singular solutions with the pole in those orthogonality relations) to obtain a contradiction when the pole converges to the boundary of one of the domains Q_j. To show that some integrals in these relations are bounded while one of them is not, we will use estimates of integrals of singular solutions given in section 3, which is the most technically difficult part of the paper.

2. Orthogonality relations. In this section, we assume that the conditions of Theorem 1.1 are satisfied and obtain some auxiliary relations which will be used in its proof.

Denote by Q_{3t} the connected component of the open set $\Omega\backslash(\overline{Q}_{1t} \cup \overline{Q}_{2t})$ whose boundary contains Γ_0. Here $Q_{j\theta}$ is $Q_j \cap \{t = \theta\}$, $j = 1, 2$. Let $Q_3 = \cup\, Q_{3t}$ over $0 < t < T$ and let $Q_4 = Q\backslash\overline{Q}_3$.

Lemma 2.1.

$$\int_{Q_1} b_1 \nabla v_1 \cdot \nabla u_2^* \, dx\, dt = \int_{Q_2} b_2 \nabla v_1 \cdot \nabla u_2^* \, dx\, dt \tag{2.1}$$

for all solutions v_1 to equation (1.1) ($j = 1$) *near $\bar{Q}_4$ that are* 0 *when $t < 0$ and solutions u_2^* to the adjoint equation $(u_2^*)_t +$* $\operatorname{div}(a_2\nabla u_2^*) = 0$ *near $\bar{Q}_4$ that are* 0 *when $t > T$.*

Proof. From well-known results about regularity of solutions to the parabolic initial-boundary value problem (1.1)–(1.3), it follows that u_j is in $C^{2,1}(Q_3)$ and in $H^{2,1}(Q_5)$, where $Q_5 = V \times (0,T)$ and V is a vicinity of $\partial\Omega$ in Ω. Due to conditions (1.2) and (1.5), both u_1 and u_2 have the same Cauchy data on $\Gamma_0 \times (0,T)$ and satisfy the same parabolic equation in Q_3; thus from uniqueness of continuation for second-order parabolic equations (see, e.g., [7, Corollary 1.2.4]), we conclude that $u_1 = u_2$ on Q_3. Letting $u = u_2 - u_1$ and subtracting the equations (1.1) with $j = 1$ from those with $j = 2$, we get

$$\operatorname{div}((a_0 + b_2\chi(Q_2))\nabla u) - u_t = \operatorname{div}((b_1\chi(Q_1) - b_2\chi(Q_2))\nabla u_1) \quad \text{in } Q. \tag{2.2}$$

Now using the definition of a weak solution to the parabolic equation under consideration, we obtain

$$\int_Q ((a_0 + b_2\chi(Q_2))\nabla u \cdot \nabla\psi + u_t\psi) = \int_Q (b_1\chi(Q_1) - b_2\chi(Q_2))\nabla u_1 \cdot \nabla\psi \tag{2.3}$$

for any function ψ from $H_0^{1,1}(Q)$. Since u and $\chi(Q_j)$ are zero outside $\overline{Q}_4 \cap \{t < T\}$, this relation remains valid for any function ψ from $H^{1,1}(Q_6)$ (where Q_6 is an arbitrary vicinity of Q_4) that is 0 when $t > T$.

If $\psi = u_2^*$ is an $H^{1,1}(Q_6)$ solution to the adjoint equation from Lemma 2.1, then integrating the left side of (2.3) by parts with respect to t and using the definition of a weak solution to this adjoint equation with the test function u (which is zero outside $Q_4 \cap \{t < T\}$), we conclude that the left side in (2.3) is zero. Thus we have relation (2.1) with u_1 instead of v_1.

Now by using the Runge property, we extend equality (2.1) onto all v_1 solving equation (1.1) with $j = 1$ near $\overline{Q}_4$ and satisfying the initial condition (1.3). Denote the space of such v_1 by X. It is sufficient to prove that solutions u_1 to the initial-boundary value problem (1.1)–(1.3) with $j = 1$ (for various g supported in $\Gamma_0 \times (0,T)$) approximate in $L_2(Q_4)$ any solution from X. We denote the space of solutions to (1.1)–(1.3) (with various g) by X_1. Indeed, let $v_1 \in X$. Then we can approximate it similarly by solutions from X in $L_2(Q_7)$, where Q_7 is a Lipschitz domain containing Q_4 with $\operatorname{dist}(\partial_x Q_7, Q_4) > 0$. From the well-known interior Schauder-type estimates for parabolic equations, it follows that these solutions from X_1 will approximate v_1 in $H^{1,0}(Q_4)$.

To prove L_2 approximation in view of the Hahn–Banach theorem, it is sufficient to show that if f from the dual space $L_2(Q_4)$ is orthogonal to X_1, then f is orthogonal to X.

Let Ω_0 be a bounded domain with C^2-boundary such that $\Omega \subset \Omega_0$, $\Omega \neq \Omega_0$, and $\partial\Omega \backslash \Gamma_0$ belong to $\partial\Omega_0$. Let $K(x,t;y,s)$ be the Green function to the first initial value problem for the operator $\partial_t + \operatorname{div}(a_1\nabla)$ in $Q_0 \times (0,T)$. Let f be orthogonal to X_1. The Green potential

$$U(x,t;f) = \int_{Q_4} fK(x,t;\;) \tag{2.4}$$

is equal to zero on $Q_0\backslash\overline{Q}_4$ because the function $u_1 = K(x,t;\ )$ belongs to X_1 if $(x,t) \in Q_0\backslash\overline{Q}_4$. Since supp $f \subset \overline{Q}_4$, this potential is a solution to the equation $-\mathrm{div}(a_0\nabla u) = u_t$ on $Q_0\backslash\overline{Q}_4$. The coefficient a_0 belongs to $C^1(\overline{Q}_0)$, so this equation has the property of unique continuation. Therefore, $U(\ ;f) = 0$ on $Q_0\backslash\overline{Q}_4$. Now let $v \in X$; then v is a solution to the homogeneous equation near $Q_5 \cup \partial_x Q_5$, where Q_5 is an open set with C^∞ lateral boundary and $\mathrm{dist}(\partial_x Q_5, \partial_x Q_4) > 0$. Using the representation of v by a single layer potential, we obtain

$$v(y,s) = \int_{\partial_x Q_5} gK(\ ;y,s)d\Gamma$$

for some $g \in C(\partial_x Q_5)$. By using this representation, (2.4), and Fubini's theorem, we obtain

$$\int_{Q_4} fv = \int_{\partial_x Q_5} gU(\ ;f) = 0$$

because $U(\ ;f) = 0$ on $\partial_x Q_5$. Accordingly, relation (2.1) is valid for any v_1 satisfying the conditions of Lemma 2.1.

The proof is complete.

Assume that

$$Q_1 \neq Q_2. \tag{2.5}$$

Then we may assume that Q_1 is not contained in Q_2. Hence, using condition (1.5) of Theorem 1.1 on Q_j, we conclude that there is a point $(x_0,t_0) \in \partial Q_1\backslash\overline{Q}_2$ such that $(x_0,t_0) \in \partial_x Q_3$. By considering $g = 0$ for $t < t_0$ and using the translations $t \to t - t_0$ and $x \to x - x_0$, we can reduce the general case to $t_0 = 0$ and $x_0 = 0$. We can choose a ball $B \subset \mathbb{R}^n$ centered at 0 and a cylinder $Z = B \times (0,\tau)$ such that $\overline{B} \subset \Omega$, $\overline{Z}$ does not intersect $\overline{Q}_2$, and $(\partial_x Q_1) \cap \overline{Z}$ is a Lipschitz surface. Due to well-known variants of the Whitney extension theorem, there is a $C^2(\overline{Q}_1 \cup \overline{Z})$-function a_3 that coincides with a_1 on Q_1. Extend a_3 onto $Q\backslash(\overline{Q}_1 \cup \overline{Z})$ as a_0.

Lemma 2.2. *Under the conditions of Lemma* 2.1*,*

$$\int_{Q_1} b_1\nabla u_3 \cdot \nabla u_2^* = \int_{Q_2} b_2\nabla u_3 \cdot \nabla u_2^*$$

for any solution u_3 to the equation $\mathrm{div}(a_3\nabla u_3) - (u_3)_t = 0$ *near $\overline{Q}_4$ which is* 0 *when* $t < 0$ *and for any solution u_2^* from Lemma* 2.1*.*

Proof. Consider u_3 and let Q_8 be an open set with C^∞-boundary $\partial_x Q_8$ and that contains Q_4 with $\mathrm{dist}(\partial_x Q_8, Q_4) > 0$ such that u_3 is a solution to the equation $\mathrm{div}(a_3\nabla u_3) - (u_3)_t = 0$ near $\overline{Q}_8$.

Introduce a sequence of open sets Q_{4k} such that (i) $Q_{4k}\backslash Z = Q_4\backslash Z$ and (ii) the (Hausdorff) distance from ∂Q_{4k} to $\partial_x Q_4$ is less than $1/k$ and $\partial_x Q_{4k} \cap Z$ does not intersect $\overline{Q}_4$. Define a coefficient a_{3k} as a_3 on $Q_8\backslash(Q_{4k}\backslash Q_4)$ and as a_0 on $Q_{4k}\backslash Q_4$. Since $\partial Q_4 \cap Z$ is a Lipschitz surface, we have

$$\mathrm{meas}_n\{a_{3k} \neq a_3\} \to 0 \quad \text{as } k \to +\infty. \tag{2.6}$$

Let u_{3k} be solutions to the initial-boundary value problems

$$\mathrm{div}(a_{3k}\nabla u_{3k}) - (u_{3k})_t = 0 \quad \text{in } Q_8, \qquad u_{3k} = u_3 \quad \text{on } \partial_x Q_8, \qquad u_{3k} = 0 \quad \text{on } Q_8 \cap \{t = 0\}.$$

Since $u_{3k} = a_0 + \chi(Q_1)b_1$ near $\overline{Q}_1$, relation (2.1) is valid for any $u_1 = u_{3k}$. The difference $u_k = u_{3k} - u_3$ satisfies the equation

$$\operatorname{div}(a_{3k}\nabla u_k) - (u_k)_t = \operatorname{div}((a_3 - a_{3k})\nabla u_3) \quad \text{in } Q_8,$$

and $u_k = 0$ on $\partial Q_8 \cap \{t < T\}$ because u_{3k} and u_3 coincide on the lateral boundary of Q_8 and when $t = 0$. From the definition of a weak solution to this initial-boundary value problem with the test function u_k, we have

$$\int_{Q_8} a_{3k}\nabla u_k \cdot \nabla u_k + \int_{Q_8 \cap \ \{t=T\}} \frac{u_k^2}{2} = \int_{Q_8} (a_3 - a_{3k})\nabla u_3 \cdot \nabla u_k.$$

According to the assumptions, $\epsilon < a_{3k}$ for certain positive ϵ. Using this inequality, dropping the second integral in the left side, and bounding the right side by the inequality $x \cdot y \leq \epsilon^{-1}/2|x|^2 + \epsilon/2|y|^2$, we obtain

$$\int_{Q_8} \epsilon|\nabla u_k|^2 \leq C(\epsilon)\int_{Q_8} |a_3 - a_{3k}|^2|\nabla u_3|^2 + \frac{\epsilon}{2}\int_{Q_8} |\nabla u_k|^2.$$

Since ∇u_3 belongs to $L_2(Q_8)$, we conclude from (2.6) that the first integral in the right side tends to 0. Therefore, ∇u_k converges to 0 in $L_2(Q_8)$. Putting $u_1 = u_{3k} = u_3 + u_k$ into relation (2.1) and letting $k \to \infty$, we complete the proof of Lemma 2.2.

3. Estimates of integrals of singular solutions. We will make use of solutions u_3 and u_2^* with singularities outside Q_4. Solutions of elliptic equations of second order with arbitrary power singularities were constructed by Alessandrini [1]; we do not know of similar results for parabolic equations. To simplify obtaining bounds on the integrals of such solutions, we introduce new variables. We can assume that the direction e_n of the x_n-axis coincides with the interior unit normal to $\partial_x Q_1 \cap \{t = 0\}$. According to our assumptions, $\partial_x Q_1$ near the origin is the graph of a Lipschitz function $x_n = q_1(x_1, \ldots, x_{n-1}, t)$ which can be assumed to be defined and Lipschitz on the whole $\mathbb{R}^n$. The substitution

$$x_k = x_k^*, \quad k = 1, \ldots, n-1, \qquad x_n = x_n^* + q_1(x_1^*, \ldots, x_{n-1}^*, t), \quad t = t^*$$

transforms the equations (1.1) into similar equations with additional first-order differentiation with respect to x_n^* multiplied by a Lipschitz function of t. The domains Q_j are transformed onto domains with similar properties and with the additional property that the points $(0, t), 0 < t < T$, belong to $\partial_x Q_1$. Since the (hyper)plane $\{x_n^* = 0\}$ is tangent to this surface at the origin, we can find a cone $\mathcal{C} = \{|x^*/|x^*| - e_n| < \theta$, $|x^*| < \epsilon\}$ such that the cylinder $\mathcal{C} \times (0, T)$ is inside Q_1. Henceforth, we drop the sign $*$.

Let K^+ be the fundamental solution of the Cauchy problem for the forward parabolic equation $\operatorname{div}(a_3\nabla u_3) - (u_3)_t = 0$ in $*$-coordinates. Let K^- be the fundamental solution of the backward Cauchy problem for the backward parabolic equation $\operatorname{div}(a_2\nabla u_2) + u_{2t} = 0$ in these coordinates. It is known that

$$K^+ = K_1^+ + K_0^+, \qquad K^- = K_1^- + K_0^-, \tag{3.1}$$

where K_1^+ and K_1^- are the principal parts of K^+ and K^- (parametrices) and K_0^+

and K_0^- are the remainders. The principal parts are

$$K_1^+(x,t;y,\tau) = \frac{C}{(a_3(y)(t-\tau))^{n/2}} \exp\left(-\frac{|x-y|^2}{4a_3(y)(t-\tau)}\right), \tag{3.2}$$
$$K_1^-(x,t;y,\tau) = \frac{C}{(a_0(y)(\tau-t))^{n/2}} \exp\left(-\frac{|x-y|^2}{4a_0(y)(\tau-t)}\right).$$

From the known bounds of fundamental solutions of parabolic equations [9, p. 377], we have

$$|\nabla_x K_0^+(x,t;y,\tau)| \le C(t-\tau)^{-n/2} \exp\left(-\frac{|x-y|^2}{(C(t-\tau))}\right), \tag{3.3}$$
$$|\nabla_x K_0^-(x,t;y,\tau)| \le C(\tau-t)^{-n/2} \exp\left(-\frac{|x-y|^2}{(C(\tau-t))}\right).$$

When $(y,0)$ and (y,τ) are outside $\overline{Q}_1$, the functions $K^+(\ ;y,0)$ and $K^-(\ ;y,\tau)$ are (x,t)-solutions to the homogeneous parabolic equations with bounded measurable coefficients satisfying zero initial and final conditions. Using Lemma 2.2 with $u_3 = K^+(\ ;y,0)$ and $u_2^* = K^-(\ ;y,\tau)$, we get

$$\begin{aligned} &\int_{Q_1\cap Z} b_1 \nabla_x K^+(\ ;y,0)\cdot\nabla_x K^-(\ ;y,\tau) \\ &\quad = -\int_{Q_1\setminus Z} b_1 \nabla_x K^+(\ ;y,0)\cdot\nabla_x K^-(\ ;y,\tau) \\ &\quad\quad + \int_{Q_2} b_2 \nabla_x K^+(\ ;y,0)\cdot\nabla_x K^-(\ ;y,\tau). \end{aligned} \tag{3.4}$$

From the estimates in (3.3) and similar estimates for $\nabla_x K_1^+$ and $\nabla_x K_1^-$, we conclude that the integrands are bounded by an integrable function uniformly with respect to y outside Q_1. By the Lebesgue dominated-convergence theorem, we may let $y \to 0$ and replace y in (3.4) by 0. Using representation (3.1), we obtain from (3.4) that

$$|I_1| \le |I_2| + |I_3|, \tag{3.5}$$

where

$$I_1 = \int_{Q_1\cap Z} b_1 \nabla_x K_1^+(\ ;0,0)\cdot\nabla_x K_1^-(\ ;0,\tau)$$

is formed from the principal parts of K and the remainders are collected in

$$I_2 = -\int_{Q_1\setminus Z} b_1 \nabla_x K^+(\ ;0,0)\cdot\nabla K^-(\ ;0,\tau) + \int_{Q_2} b_2 \nabla_x K^+(\ ;0,0)\cdot\nabla_x K^-(\ ;0,\tau)$$

and

$$\begin{aligned} I_3 = \int_{Q_1\cap Z} b_1 (&\nabla_x K_1^+(\ ;0,0)\cdot\nabla K_0^-(\ ;0,\tau) + \nabla_x K_0^+(\ ;0,0)\cdot\nabla_x K_1^-(\ ;0,\tau) \\ &+ \nabla_x K_0^+(\ ;0,0)\cdot\nabla K_0^-(\ ;y,\tau)). \end{aligned}$$

In the following three lemmas, I_1 is bounded from below and I_2 and I_3 is bounded from above.

LEMMA 3.1.

$$|I_1| \geq C^{-1}\tau^{-n}\int_0^{\epsilon}\rho^{n-1}e^{-4p^2/(m\tau)}d\rho,$$

where $m = \inf(a_3, a_0)$ *over* Q.

Proof. Using the fact that $b_1(0) \neq 0$ and choosing ϵ in the definition of $\mathcal{C}$ to be sufficiently small, we obtain

$$\begin{aligned}|I_1| \geq\ & C^{-1}\int_{\mathcal{C}\times(0,\tau)}\nabla_x K_1^+(x,t;0,0)\cdot\nabla_x K_1^-(x,t;0,\tau)\\ =\ & C^{-1}\int_0^{\tau}\int_{\mathcal{C}}t^{-n/2-1}\exp\left(-\frac{|x|^2}{a_3(x)t}\right)x\cdot(\tau-t)^{-n/2-1}\exp\left(\frac{|x|^2}{a_0(x)(\tau-t)}\right)x\,dx\,dt\\ \geq\ & C^{-1}\int_{\mathcal{C}}\int_0^{\tau/2}|x|^2((\tau-t)t)^{-n/2-1}\exp\left(-\frac{|x|^2\tau}{mt(\tau-t)}\right)dt\,dx.\end{aligned}$$

Using the inequality

$$\frac{1}{t\tau} \leq \frac{1}{t(\tau-t)} \leq \frac{2}{t\tau} \quad \text{when } 0 < t < \frac{\tau}{2}, \tag{3.6}$$

we bound from below the integral shown above by

$$\begin{aligned}C^{-1}&\int_{\mathcal{C}}\int_0^{\tau/2}|x|^2\frac{1}{(t\tau)^{n/2+1}}\exp\left(-\frac{2|x|^2}{mt}\right)dt\,dx\\ &= \frac{1}{C\tau^{n/2+1}}\int_{\mathcal{C}}|x|^{2-n}\int_{\frac{4|x|^2}{m\tau}}^{\infty}w^{n/2-1}e^{-w}dw\,dx,\end{aligned}$$

where we substituted $w = 2|x|^2/mt$.

The function $w^{n/2-1}$ is increasing, so replacing it by its minimal value at $w = 4|x|^2/(m\tau)$, we bound the last integral from below by

$$\int_{\mathcal{C}}\tau^{1-n/2}\left(\int_{(4|x|^2/(m\tau),\infty)}e^{-w}dw\right)dx = C^{-1}\tau^{1-n/2}\int_{(0,\epsilon)}\rho^{n-1}e^{-4\rho^2/(m\tau)}d\rho.$$

The proof is complete.

LEMMA 3.2.

$$|I_2| \leq C\tau^{-n/2+1}\epsilon^{-2}e^{-\epsilon^2/(M\tau)},$$

where M *depends only on* $\sup(a_3, a_0)$ *over* Q.

Proof. I_2 consists of two integrals. The first one is bounded by

$$\begin{aligned}C&\int_{\epsilon<|x|<R,0<t<\tau}|\nabla_x K^+(\ ;0,0)\cdot\nabla_x K^-(\ ;0,\tau)|\\ \leq\ & C\int_{\epsilon<|x|<R}\int_0^{\tau/2}\frac{1}{((\tau-t)t)^{n/2+1/2}}\exp\left(-\frac{|x|^2\tau}{Mt(\tau-t)}\right)dt\,dx.\end{aligned}$$

The bound on $|\nabla_x K^+ \cdot \nabla_x K^-|$ follows from the direct differentiation of (3.2), the inequality

$$|x|(t-\tau)^{-n/2-1} \exp\left(-\frac{|x|^2}{4(t-\tau)}\right) \leq C(t-\tau)^{-n/2-1/2} \exp\left(-\frac{|x|^2}{8(t-\tau)}\right),$$

and the bounds in (3.3).

Applying inequality (3.6) as above, we bound the last integral by

$$\frac{C}{\tau^{n/2+1/2}} \int_{\epsilon<|x|<R} \int_0^{\tau/2} \frac{1}{t^{n/2+1/2}} \exp\left(-\frac{|x|^2}{Mt}\right) dt\, dx$$
$$\leq \frac{C}{\tau^{n/2+1/2}} \int_{\epsilon<|x|<R} |x|^{1-n} \int_{\frac{2|x|^2}{M\tau}}^{\infty} w^{n/2-1} w^{-1/2} e^{-w} dw\, dx$$

when we use the substitution $w = |x|^2/(Mt)$. The function $w^{-1/2}$ is decreasing. Replacing it by its value at $2|x|^2/(M\tau)$, we increase the integral, and we also use the inequality $w^{n/2-1}e^{-w} \leq Ce^{-w/2}$ and calculate the resulting integral with respect to w. Then the last integral will be less than

$$C\tau^{-n/2} \int_{\epsilon<|x|<R} |x|^{-n} \exp(-|x|^2/(M\tau)) dx$$
$$\leq C\tau^{-n/2} \int_{(\epsilon,\infty)} \rho^{-2} \rho \exp(-\rho^2/(M\tau)) d\rho$$

when we use the polar coordinates in $\mathbb{R}^n$. Replacing ρ^{-2} by its maximal value at ϵ and calculating the remaining integral with respect to ρ, we complete the bounding of the integral over $Q_1 \backslash Z$.

A similar argument works for the integral over Q_2.

The proof is complete.

Lemma 3.3.

$$|I_3| \leq C\epsilon\tau^{-n/2},$$

where M depends only on the upper bounds of $|a_3|$, $|a_0|$.

Proof. We bound the integral of the first of the three functions, forming I_3 as defined after (3.5).

As follows from (3.2), (3.3), and the argument in Lemma 3.2, replacing $|x|$ by some power of $(t-\tau)$, the absolute value of this integral is less than

$$C \int_{|x|<\epsilon} \int_{(0,\tau/2)} ((\tau-t)t)^{-n/2} t^{-1/2} \exp(-|x|^2\tau/(Mt(\tau-t)) dt\, dx$$
$$\leq C \int_{|x|<\epsilon} \int_{(0,\tau/2)} (\tau t)^{-n/2} t^{-1/2} \exp(-|x|^2/(Mt)) dt\, dx,$$

where we used inequality (3.6). Substituting $w = |x|^2/(Mt)$ in the inner integral yields

$$C\tau^{-n/2} \int_{|x|<\epsilon} |x|^{1-n} \int_{(2|x|^2/(M\tau),\infty)} w^{n/2-3/2} e^{-w} dw\, dx \leq C\tau^{-n/2} \int_{|x|<\epsilon} |x|^{1-n} dx.$$

Using the polar coordinates, we bound the last integral by $C\epsilon$.

The other terms can be bounded in a similar way. The proof is complete.

4. Proof of Theorem 1.1. Now we will complete the proof of Theorem 1.1. Let

$$\epsilon^2 = E\tau, \tag{4.1}$$

where (large) E will be chosen later.

First, we bound I_1 from below. From Lemma 3.1, substituting $w = 4\rho^2/(m\tau)$ in the integral and using condition (4.1), we obtain

$$|I_1| \geq C^{-1}\tau^{-n/2} \int_{(0,4E/m)} w^{n/2-1} e^{-w} dw \geq C^{-1}\tau^{-n/2} \tag{4.2}$$

provided $E > m$.

From (3.5), Lemmas 3.1–3.3, (4.1), and (4.2), it follows that

$$C^{-1}\tau^{-n/2} \leq C(\tau^{-n/2+1}\epsilon^{-2} \exp(-E/M) + \tau^{-n/2}\epsilon).$$

Using (4.1) again and multiplying both sides by $C\tau^{n/2}$, we obtain

$$1 \leq CE^{-1} \exp(-E/M) + C\epsilon \leq CE^{-1} + C\epsilon.$$

Let $\tau < 1$. Choose E so large that $E^{-1} < 1/(4C)$ and $\epsilon < 1/(4C)$; then the right side is smaller than $1/2$. We have a contradiction.

This contradiction shows that $Q_1 = Q_2$.

The next step of the proof is to show that

$$b_1 = b_2 \quad \text{on } \partial_x Q_1. \tag{4.3}$$

As in the proof for Q_j, we assume the opposite. Then we can assume that the origin $0 \in \partial_x Q_1$ and $b_1(0) < b_2(0)$. By continuity, $b_1(0) - b_2(0) > C^{-1}$ for some C on a certain ball B centered at the origin. Let $Z = B \times (0, T)$. Extend a_2 from Q_2 onto $\mathbb{R}^n$ as a C^2-function $a_4 > 0$. By repeating the proof of Lemma 3.2, we obtain the following orthogonality relation:

$$\int_{Q_1} (b_1 - b_2)\nabla u_3 \cdot \nabla u_4^* = 0 \tag{4.4}$$

for all solutions u_3 to the equation $\operatorname{div}(a_3 \nabla u_3) - u_{3t} = 0$ near Q_4 which are zero when $t < 0$ and for all solutions u_4^* to the adjoint equation $\operatorname{div}(a_4 \nabla u_4) + u_{4t} = 0$ near Q_4 which are zero when $t > T$. Let K^+ be a fundamental solution to the forward Cauchy problem for the first equation and K^- be the fundamental solution to the backward Cauchy problem for the adjoint equation with the coefficient a_4. Using the representation (3.1) of these fundamental solutions and splitting Q_1 into $Q_1 \cap Z$ and its complement, as in section 3, we obtain from (4.4) the inequality

$$|I_4| \leq |I_5| + |I_6|, \tag{4.5}$$

where

$$I_4 = \int_{Q_1 \cap Z} (b_1 - b_2) \nabla_x K_1^+ \cdot \nabla_x K_1^-$$

is related to the supposedly singular part and

$$I_5 = \int_{Q_1 \setminus Z} (b_1 - b_2) \nabla K^+ \cdot \nabla K^-,$$

$$I_6 = \int_{Q_1} (b_1 - b_2) \nabla K_0^+ \cdot \nabla K_0^{-2}.$$

It is easy to see that Lemmas 3.1, 3.2, and 3.3 are valid for I_4, I_5, and I_6, respectively. Therefore, as in the proof above, we arrive at the contradiction that $Q_1 = Q_2$.

This shows that the assumption about b_1 and b_2 is wrong and that $b_1 = b_2$ on $\partial_x Q_1$.

Let Ω_0 be the intersection of all $Q_{1\theta}$ over $0 < \theta < T$. Since b_1 and b_2 do not depend on t and are equal on $\partial_x Q_1$, they coincide on $Q_{1\theta} \setminus \Omega_0$. Letting $Q_0 = \Omega \times (0, T)$, we obtain from (4.4) the relation

$$\int_{Q_0} (b_1 - b_2) \nabla u_3 \cdot \nabla u_4^* = 0$$

for all u_3 and u_4^* in (4.4). As in the proof of Lemma 3.2, this implies that

$$\int_{Q_0} (b_1 - b_2) \nabla u_6 \cdot \nabla u_6^* = 0 \tag{4.6}$$

for solutions u_5 to the equation $\operatorname{div}((a_0 + b_1 \chi(Q_0)) \nabla u_5) - u_{5t} = 0$ near Q_0 which are zero when $t < 0$ and for solutions to the adjoint equation $\operatorname{div}((a_0 + b_2 \chi(Q_0)) \nabla u_6^*) - u_{6t}^* = 0$ near Q_0 which are zero when $t < T$.

Observe that by choosing T small, we can guarantee that Ω_0 is a Lipschitz domain. Indeed, for any point of $\partial_x Q_1 \cap \{t = 0\}$, there is a neighborhood where Q_1 is the subgraph of the Lipschitz function $x_j < q_j(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n, t)$. We can cover the compact set $\partial_x Q_1 \cap \{t = 0\}$ by a finite number of such neighborhoods. Then there is T_1 such that $\partial_x Q_1 \cap \{t < T_1\}$ is contained in the union of these neighborhoods. Let $T = T_1$; then Ω_0 is Lipschitz because locally (in the corresponding neighborhood) its boundary is given by the equation $x_j = \inf q_j(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n, t)$ over $t \in (0, T)$, and the inf of a family of uniformly Lipschitz functions is a Lipschitz function.

Now we will show that the equations for u_5 and u_6 have the same lateral Dirichlet-to-Neumann maps. Let u_5 and u_6 be a solution to these equations with zero initial conditions and the same lateral Dirichlet data. By subtracting these equations and letting $u = u_0 - u_5$, we obtain

$$\operatorname{div}((a_0 + b_2 \chi(Q_0)) \nabla u) = \operatorname{div}((b_1 - b_2) \chi(Q_0) \nabla u_5) \quad \text{in } Q.$$

From the definition of a weak solution of this equation, we have

$$\int_{\partial\Omega \times (0,T)} a_0 u_\nu \psi - \int_Q ((a_0 + b_2 \chi(Q_0)) \nabla u \cdot \nabla \psi - \int_Q u_t \psi = - \int_{Q_0} (b_1 - b_2) \nabla u_5 \cdot \nabla \psi$$

for any function $\psi \in H^{1,1}(Q)$. Using $\psi = u_6^*$, integrating by parts in the third integral of the left side, and again using the definition of a weak solution to the equation $\operatorname{div}((a_0 + b_2 \chi(Q_0)) \nabla u_6^*) + u_{6t}^* = 0$ with the test function u which is zero on

$\partial Q \cap \{t < T\}$, we conclude that the sum of the second and third integrals in the left side is zero. The right side is zero due to (4.6). Thus the first integral in the left side is zero. Since the lateral Dirichlet data $\psi = u_6^*$ can be any function in $C_0^\infty(\partial\Omega \times (0,T))$, we get $u_\nu = 0$ on $\partial\Omega \times (0,T)$. Therefore, $u_{5\nu} = u_{6\nu}$ on the lateral boundary, which means that we have the same lateral Dirichlet-to-Neumann maps.

Take as the Dirichlet data g a function which does not depend on t when $t > \tau$. Since the coefficients of the equations $\operatorname{div}((a_0 + b_j\chi(Q_0)\nabla u_j) - u_{jt} = 0$ are time independent, the solution $u_j(x,t)$ of the initial-boundary value problems on $\Omega \times (0,\infty)$ will be analytic with respect to $t > \tau$. They have the same Cauchy data on $\partial\Omega \times (0,T)$; therefore, as above, by uniqueness in the lateral Cauchy problem, $u_5 = u_6$ on $(\Omega\backslash\Omega_0) \times (0,T)$. By uniqueness of the analytic continuation, they are equal also on $(\Omega\backslash\Omega_0) \times (0,\infty)$. Now we modify the argument of [6] and consider the Laplace transforms

$$U_j(x,s) = \int_{(0,\infty)} e^{-st} u_j(x,t)\, dt.$$

They solve the following Dirichlet problems:

$$\operatorname{div}((a_0 + b_j\chi(\Omega_0))\nabla U_j) - sU_j = 0 \quad \text{in } \Omega, \qquad U_j = G \quad \text{on } \partial\Omega, \tag{4.7}$$

and $U_5 = U_6$ on $\Omega\backslash\Omega_0$. Letting $\tau \to 0$ we obtain $G(x,s) = g_0(x)s^{-1}$, where $g_0(x) = g(x,t)$ when $t > \tau$. Applying the results of [5] and [11] on identification of elliptic equations, we conclude that $b_1 = b_2$ on Ω_0. In fact, this result is obtained in [5] when $n \geq 3$, but the recent global uniqueness theorem of Nachman [10] extends it to $n = 2$.

The proof is complete.

REFERENCES

[1] G. Alessandrini, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–273.

[2] H. Bellout, *Stability result for the inverse transmissivity problem*, J. Math. Anal. Appl., 168 (1992), pp. 13–27.

[3] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[4] A. Friedman and V. Isakov, *On uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 553–580.

[5] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[6] V. Isakov, *Uniqueness for inverse parabolic problems with a lateral overdetermination*, Comm. Partial Differential Equations, 14 (1989), pp. 681–689.

[7] V. Isakov, *Inverse Source Problems*, Math. Surveys and Monographs Series, Vol. 34, AMS, Providence, RI, 1990.

[8] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II: *Interior results*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[9] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs 23, AMS, Providence, RI, 1968.

[10] A. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. Math., 142 (1995), pp. 71–96.

[11] J. Sylvester and G. Uhlmann, *Global uniqueness theorem for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.

[12] W. Yeh, *Review of parametric identification procedures in ground water hydrology*, Water Resource Res., 22 (1986), pp. 95–108.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 60–93, January 1997

THE EXISTENCE OF TRAVELLING WAVE SOLUTIONS OF A GENERALIZED PHASE-FIELD MODEL[*]

P. W. BATES[†], P. C. FIFE[‡], R. A. GARDNER[§], AND C. K. R. T. JONES[¶]

Abstract. This paper establishes the existence and, in certain cases, the uniqueness of travelling wave solutions of both second-order and higher-order phase-field systems. These solutions describe the propagation of planar solidification fronts into a hypercooled liquid. The equations are scaled in the usual way so that the relaxation time is $\alpha\varepsilon^2$, where ε is a nondimensional measure of the interfacial thickness. The equations for the transition layer separating the two phases form a system identical to that for the travelling-wave problem, in which the temperature is strongly coupled with the order parameter. Thus there is no longer a well-defined temperature at the inteface, as is the case in the more frequently studied situation in which the liquid phase is undercooled but not hypercooled.

For phase-field systems of two second-order equations, we prove a general existence theorem based upon topological methods. A second, constructive proof based upon invariant-manifold methods is also given when the parameter α is either sufficiently small or sufficiently large. In either regime, it is also proved that the wave and the wave velocity are globally unique.

Analogous results are also obtained for generalized phase-field systems in which the order parameter solves a higher-order differential equation. In this paper, the higher-order tems occur as a singular peturbation of the standard (isotropic) second-order equation. The higher-order terms are useful in modelling anisotropic interfacial motion.

Key words. travelling waves, phase-field equations, hypercooling

AMS subject classification. 35K55

PII. S0036141095283820

1. Introduction.

The phase-field system,

$$\begin{aligned}(u + \lambda W(\varphi))_t &= \nabla^2 u,\\ \alpha\varepsilon^2\varphi_t &= \varepsilon^2\Lambda\varphi + F(\varphi, u),\end{aligned}$$

is a well-established model for describing the behavior of phase fronts in materials that are undergoing a transition between the liquid and solid phase (see, e.g., [2, 9, 8, 3, 4]). Here u is nondimensional temperature, φ is an order parameter, $u + \lambda W(\varphi)$ is the energy density, Λ is an elliptic partial differential operator, ε is a length scale associated with the interfacial thickness and also serves as a nondimensional surface tension, and α is $O(1)$ with respect to ε. The functions W and F are characterized precisely in the next section; however, we note here that the assumptions about W include the cases in which W is either a linear or a quadratic function and that F is a bistable function. The nonnegative parameter λ is included for convenience in the analysis to

[*] Received by the editors March 20, 1995; accepted for publication (in revised form) November 3, 1995.

http://www.siam.org/journals/sima/28-1/28382.html

[†] Department of Mathematics, Brigham Young University, Provo, UT 84602 (peter@math.byu.edu). The research of this author was partially supported by NSF grant DMS-9305044 and the Isaac Newton Institute, Cambridge.

[‡] Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (fife@math.utah.edu). The research of this author was partially supported by NSF grant DMS-9201714 and the Isaac Newton Institute, Cambridge.

[§] Department of Mathematics, University of Massachusetts, Amherst, MA 01003 (gardner@math.umass.edu). The research of this author was partially supported by NSF grants DMS-898922384 and DMS-9300848-001.

[¶] Division of Applied Mathematics, Brown University, Providence, RI 02912 (ckrtj@cfm.brown.edu). The research of this author was partially supported by NSF grant DMS-9402774.

follow. In [1], the term λW is denoted simply by w. One advantage of phase-field equations over sharp interface models is that they incorporate both equations for the field variables away from the interface and the proper interface conditions within a single system. In addition, they are computationally advantageous.

We shall employ the usual measure of degree of undercooling $\Delta = (\hat{c}/\ell)(T_m - T_-)$, where $\hat{c}$, ℓ, T_m, and T_- are the specific heat, latent heat at $T = T_-$, melting temperature, and temperature of the undercooled melt. In dimensionless variables, the value $\Delta = 1$ is the threshold beyond which the melt is hypercooled (see the discussion following Hypothesis 2 in section 2). A frequently studied problem is that of describing the advance of a solid into a supercooled liquid for which $\Delta < 1$ (see, e.g., [4, 8]). Then at locations away from the interface, the temperature of the solid being formed is approximately the melting temperature T_m, and there can be no planar front moving with constant velocity (it will decelerate). The analogous problem for a hypercooled medium, in which $\Delta > 1$, has received less attention. In this case, the solidification front advances more rapidly into the liquid and can have constant speed. The existence of these travelling fronts within the framework of the phase field models is the principal aim of this paper.

The asymptotic analysis of the phase-field equations in this regime is performed in a companion paper [1], in which matched asymptotic expansions of the solution for small ε are constructed. In contrast to the case where $\Delta < 1$, it turns out that with hypercooling, both the order parameter *and* the temperature are discontinuous at the transition layer (phase interface) in the limit as $\epsilon \to 0$. This implies that the "inner" equations within the transition layer will be significantly more difficult to solve than in the case where $\Delta < 1$ since the travelling wave equations which give the fine structure of the solution in the interface now consist of a system in which the temperature equation is coupled with the equation for the order parameter. When $\Delta < 1$, temperature is constant to lowest order in the transition layer, and the resulting connection problem is a scalar ODE for φ. In this connection, it should be noted that the existence of travelling waves in the case where $\Delta > 1$ was proved by Caginalp and Nishiura [5] for a variant of the phase-field equations, used in other papers as well, in which the function F takes the form $F(\varphi, u) = F_o(\varphi) + \varepsilon u$. In the connection problem for this model, as in the present paper, u is not almost constant in the layer so that the travelling wave problem consists of a system rather than a scalar equation. However, the problem is simplified considerably by the fact that the second equation in the system (which contains the function F) is coupled only weakly to the variable u, allowing the use of a perturbative argument. In this variant model, ε has a different meaning and, most importantly, different physical assumptions are made. See [1] for a comparison of the two versions.

We shall prove several theorems on the existence and the uniqueness of travelling plane wave solutions to the phase-field system in the hypercooled regime. As shown in [1], the system for plane waves is precisely the set of "inner" equations for the transition layer.

The first model we consider is the standard isotropic phase-field system in one space variable x, in which $\Lambda = \partial^2/\partial x^2$. If $\xi = x/\varepsilon - ct/\varepsilon^2$ and u and φ are solutions of the PDEs that depend only upon ξ, then u and φ satisfy the travelling wave ODEs

$$-c(\dot{u} + \lambda \dot{W}(\varphi)) = \ddot{u},$$
$$-c\alpha\dot{\varphi} = \ddot{\varphi} + F(\varphi, u).$$

We first formulate and prove a theorem, Theorem 2.1, on the existence of heteroclinic solutions for this system using the Conley index. The theorem is quite general, requiring only geometric constraints on the nonlinearities in the equations (see Hypotheses 1–5 in section 2 below). The price for such generality is that the proof supplies relatively little information about the qualitative properties of the profile and the wave speed. It is not difficult to see that the components of the profile can exhibit a rich variety of different qualitative forms, depending on the parameters and the nonlinearities in the equations. In section 6, we present numerical results giving some of the wave profiles that can be observed.

In the asymptotics in [1], it is important to have more information about the wave and the associated wave velocity c. In particular, it is important to know that the wave speed is unique and that it depends smoothly on parameters in the equations. In section 4, we give a second proof of the existence of the profile when the parameter α is either small or large under weaker geometric constaints than those required in Theorem 2.1. The proof, which is constructive and relies upon invariant-manifold machinery, provides the local uniqueness of the wave and the wave velocity as well as their differentiable dependence on parameters. If, in addition, the more stringent constraints of Theorem 2.1 are satisfied, it is shown that the wave and the wave velocity are globally unique relative to all possible wave solutions in the isolating region constructed in section 3 and all negative values of the wave-velocity parameter.

In section 5, we consider phase-field equations where the operator Λ is a higher-order differential operator. Equations of this type are obtained by retaining higher-order terms of up to some arbitrarily prescribed order in the Ginzburg–Landau formalism used to express the self-interaction term in the free-energy functional as a differential operator. The retention of higher-order terms provides a natural way of introducing anisotropy into the equations; see [1] for a detailed derivation. The resulting travelling wave system for the inner equations at a point on the interface is

$$\begin{aligned} -c(\dot{u} + \lambda \dot{W}(\varphi)) &= \ddot{u}, \\ -c\alpha\dot{\varphi} &= \Lambda(\theta)\varphi + F(\varphi, u), \end{aligned}$$

where the vector parameter θ consists of the angles that the outward normal to the interface at this point makes with the coordinate axes (see [1]). Specifically, the operator $\Lambda(\theta)$ has the form

$$\Lambda(\theta) = \sum_{i=1}^{m} \mu^{2i-2} b_i(\theta) d^{2i}/d\xi^{2i},$$

where m is odd and the coefficients $b_i(\theta)$ are smooth positive functions of θ. In [1], it is shown that if the interaction function in the free-energy functional possesses anisotropy of order $2m$, then terms in the differential equations of order at least $2m$ must be retained in order to observe the required degree of anisotropy in solutions of the truncated equations. Also, we remark that m should be odd if Λ is to be elliptic.

The parameter μ has physical significance as the ratio of two microscopic characteristic lengths [1]. We assume that μ is small enough since the existence proof entails perturbing off the case where $\mu = 0$. This generalizes earlier work on the existence of travelling waves for the scalar problem (wherein u is constant) in the sixth-order case [11]. The results in [11] are generalized to scalar equations of arbitrarily high order in [1]. The scalar equation arises as the inner $O(1)$ approximation when the relaxation time for the phase variable, which we have denoted by $\alpha\varepsilon^2$, is assumed to be $O(\varepsilon)$ rather than the $O(\varepsilon^2)$.

In section 5, we obtain a global existence result similar to Theorem 2.1 for the higher-order connection problem under the assumption that μ is sufficiently small. We also give a constructive proof which again implies the uniqueness of the wave speed c and its differentiable dependence $c = c(\theta)$ on the parameter θ under the additional assumption that α is either small or large. This last result is needed in [1] in the asymptotic description of anisotropic interfacial motion in hypercooled solidification. In particular, the asymptotics in [1] lead to a Hamilton–Jacobi-type equation for the interface. The Hamiltonian is essentially the wave speed c of the above connection problem as a function $c = c(\theta, s)$ of the angles θ and, in the event that the initial data are in the two phases, position s along the interface.

The proofs of some of the main theorems are somewhat lengthy and technical. Readers seeking an overview of the results obtained here should first read section 2, wherein the basic structural hypotheses are formulated, and then proceed to the statement of Theorems 4.2 and 4.3 in section 4, wherein the two asymptotic limits of the wave for small and large α are described. Further qualititative understanding of the structure of various waves that appear can be obtained from the numerical results presented in section 6.

2. A global result. In this section we formulate the general travelling plane wave problem for the second-order system described in the introduction. Take the far-field liquid temperature $u_- < 0$ and the phase variable there, φ_-, to be such that the state (φ_-, u_-) lies on the equilibrium curve $F(\varphi, u) = 0$ on the right-hand ascending branch (see Figure 1). Here the right equilibrium branch is associated with the liquid phase. Note that with this convention, large φ is associated with the liquid phase while small φ along the left branch of equlibria is associated with the solid phase. Hence here φ is more appropriately thought of as a disorder parameter rather than an order parameter.

A solution of the second-order travelling wave system which is asymptotic to the state (φ_-, u_-) at $\xi = -\infty$ satisfies the system of ODEs

$$\begin{aligned} \dot{u} &= -cg(\varphi, u), \\ \dot{\varphi} &= \psi, \\ \dot{\psi} &= -c\alpha\psi - F(\varphi, u), \end{aligned} \tag{1}$$

where $c < 0$ is the wave speed, $\alpha > 0$, and

$$g(\varphi, u) = u - u_- + \lambda(W(\varphi) - W(\varphi_-)).$$

$W(\varphi)$ is a function which is either monotone decreasing or concave (see [8]). In particular, we could assume for simplicity that $W(\varphi) = A\varphi - B\varphi^2$ with $A, B \geq 0$. Additional hypotheses concerning W are formulated below.

We shall consider functions F and W which, in addition to the above, satisfy the following hypotheses.

HYPOTHESIS 1.

1. *F is a C^1 function which is "bistable" in φ for each fixed u in some interval $-u_m < u < u_m$, i.e., for each such u, $F(\varphi, u)$ has precisely three roots,*

$$h_\ell(u) < h_*(u) < h_r(u).$$

2. *$\int_{h_\ell(u)}^{h_r(u)} F(\varphi, u)\, d\varphi$ has the same sign as u and vanishes if and only if $u = 0$.*
3. *$\frac{\partial F}{\partial \varphi}(h_{\ell,r}(u), u) < 0$ for $-u_m < u < u_m$, and $\frac{\partial F}{\partial u}(\varphi, u) > 0$.*

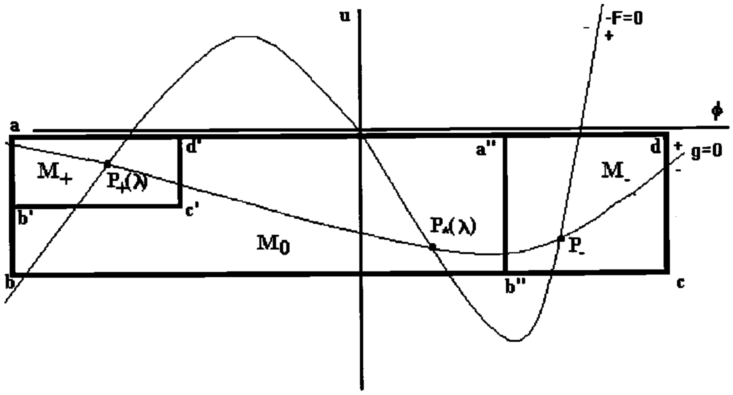


FIG. 1. *Null sets of F and g, and the boxes M_0 and $M_\pm$.*

For example, the function $F = \varphi - \varphi^3 + u$ is easily seen to satisfy each of the conditions in Hypothesis 1, as does the function $F = (1-\varphi)(\varphi + a(u))(\varphi+1)$, where $a(u)$ is a monotone increasing function with $a(0) = 0$. We remark that these conditions include the class of thermodynamically consistent phase-field equations of Fife and Penrose [8]. Since the far field at $-\infty$ is assumed to be in the liquid phase, it is appropriate to consider only waves with negative speeds c since hypercooling should produce solidification fronts propagating from the solid into the liquid. The above hypothesis is consistent with this physical requirement.

The parameter u_- is regarded as a free parameter in the range $-u_m < u_- < 0$ and the parameter φ_- is then determined by the equation $\varphi_- = h_r(u_-)$. The point $P_- = (u_-, \varphi_-, 0)$ is therefore a rest point of (1) which lies along the right branch of $F = 0$. We next impose some additional hypotheses on the null sets $F = 0$ and $g = 0$ that are needed to ensure the existence of the wave. These hypotheses can be seen to be fulfilled through a judicious choice of the remaining parameter λ. In the following, λ is allowed to vary over some interval of the form $0 \le \lambda \le \lambda_0$.

HYPOTHESIS 2 (hypercooling). *The null clines $F = 0$ and $g = 0$ of the nonlinearities in* (1) *are as depicted in Figure* 1. *More precisely, there exists $\lambda_0 > 0$ such that for $\lambda \in [0, \lambda_0]$, the equations in* (1) *admit exactly three rest points,*

$$\begin{aligned} P_- &= (u_-, \varphi_-, 0), \\ P_*(\lambda) &= (u_*(\lambda), \varphi_*(\lambda), 0), \\ P_+(\lambda) &= (u_+(\lambda), \varphi_+(\lambda), 0), \end{aligned}$$

where $\varphi_+(\lambda) = h_\ell(u_+(\lambda))$, $\varphi_- = h_r(u_-)$, and $\varphi_*(\lambda) = h_*(u_*(\lambda))$. Furthermore, we assume that

$$u_- < u_+ < 0.$$

It follows from the above that the coordinates of $P_*(\lambda)$ and $P_\pm(\lambda)$ are implicit functions of the temperature u_-. Hence the only undetermined parameter is the wave speed c.

From Figure 1, it can be seen that it is possible for the temperature $u_+(\lambda)$ to be positive at the rest point $P_+(\lambda)$ associated with the solid phase. This is clearly a nonphysical state. The temperature u_s of the solid phase should therefore be defined as $u_s = \min\{u_+(\lambda), 0\}$. In the more typical solidification process (supercooling), the temperature u_s of the solid at the interface is zero. The last condition in Hypothesis 2 requires that the temperature in the solid phase be strictly negative, which can be taken as the definition of hypercooling.

It is easily seen that hypercooling can also be defined in terms of the parameter Δ mentioned in the introduction. In our dimensionless variables, the latent heat is

$$\ell = \lambda(W(\varphi_-) - W(\varphi_s)) > 0 \quad (\varphi_s = h_\ell(u_s))$$

so that the parameter Δ is $-u_-/\ell$. It then follows that $u_+/\ell = -\Delta + 1$ so that hypercooled solidification occurs when $\Delta > 1$ and the transition to hypercooling occurs when $\Delta = 1$. The problem in which Δ is near unity is interesting. Some numerical experiments in this parameter range are presented in section 6.

With a slight abuse of notation we shall also use $P_\pm$ and P_* to denote the points $(\varphi_\pm, u_\pm)$ and (φ_*, u_*) in the (φ, u) plane, and we shall usually suppress their dependence on λ. We remark that the (physical) condition $u_- < u_+ < 0$ in Hypothesis 2 implicitly places some constraints on W and λ.

Hypothesis 3. *There exists a rectangle M_0 in the (φ, u) plane with edges parallel to the coordinate axes and with vertices at $a, b, c,$ and d as depicted in Figure* 1 *such that for each $\lambda \in [0, \lambda_0]$,*

$$\begin{aligned} -F < 0 \quad &\textit{along } \bar{ab}; \quad -F > 0 \quad \textit{along } \bar{cd}; \\ g < 0 \quad &\textit{along } \bar{bc}; \quad\;\;\; g > 0 \quad \textit{along } \bar{ad}. \end{aligned}$$

Hypothesis 4. *For each $\lambda \in [0, \lambda_0]$, there exist nested families of rectangles $M_\pm(\tau)$, $0 \le \tau \le 1$, such that* (i) $M_\pm(0) = P_\pm$, (ii) *$M_\pm(\sigma) \subset M_\pm(\tau)$ for $\sigma \le \tau$, and* (iii) *$M_\pm(\tau)$ has edges parallel to the coordinate axes and such that for $\tau > 0$, the null cline $g = 0$ intersects $\partial M_\pm(\tau)$ only along the vertical edges and the null cline $F = 0$ intersects $\partial M_\pm(\tau)$ only along the horizontal edges. Furthermore, if $M_\pm = M_\pm(1)$, then the vertices a, b', c', and d' of M_+ and the vertices a'', b'', c, and d of M_- are as depicted in Figure* 1.

It is easily seen that Hypotheses 2–4 will be satisfied whenever λ_0 is sufficiently close to zero since in that case, the rectangles M_0 and $M_\pm$ can be chosen to be thin in the vertical dimension.

Next, let (φ_w, u_w) be the point along $\{g = 0\} \cap M_0$ for which u is minimal. It follows from (i) and (ii) of Hypothesis 1 that if $r(\varphi) = -F(\varphi, u_w)$, then $r(\varphi)$ has three distinct roots at

$$\hat{\varphi}_- = h_r(u_w), \qquad \hat{\varphi}_* = h_*(u_w), \qquad \hat{\varphi}_+ = h_\ell(u_w).$$

Let $R'(\varphi) = r(\varphi)$ with $R(\hat{\varphi}_+) = 0$. It follows from Hypothesis 1 that $R(\varphi)$ is as depicted in Figure 2.

Hypothesis 5. *With λ_0 set in accordance with the previous hypotheses, let*

$$\begin{aligned} \varphi_r^+ &= h_r(u_+(\lambda)), \\ \varphi_*^+ &= h_*(u_+(\lambda)); \end{aligned}$$

assume for each $\lambda \in [0, \lambda_0]$ that

$$R(\varphi_*^+) > R(\varphi_r^+).$$

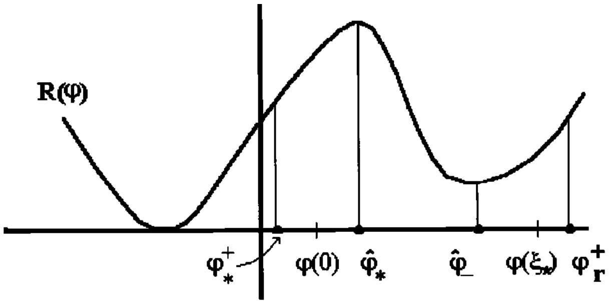


FIG. 2. *The primitive $R(\varphi)$ of $-F(\varphi, u_w)$.*

If λ_0 is near zero, $g = 0$ is nearly a horizontal line; and the rectangles M_0 and $M_\pm$ of Hypotheses 3 and 4 can be chosen to be arbitrarily thin in the vertical dimension. It then follows that $u_+(\lambda)$ is close to u_- so that $h_*(u_+(\lambda))$ will be near $\hat{\varphi}_*$ and $h_\ell(u_+(\lambda))$ will be near $\hat{\varphi}_+$. From Figure 2, we see that R has a local maximum at $\hat{\varphi}_*$ and an adjacent local minimum at $\hat{\varphi}_+$, and it therefore follows that for $\lambda \in [0, \lambda_0]$, Hypothesis 5 will always be satisfied for small λ_0.

Remark. Hypotheses 1 and 2 are natural and necessary conditions for the existence of a hypercooled phase front. It turns out that in the asymptotic regimes of either large or small α, they are also sufficient; see section 4. Hypotheses 3 and 4 are more restrictive, but they are still quite reasonable requirements for a global theory. Hypothesis 5 is somewhat more artificial. It is used below to rule out certain internal tangencies in the construction of an isolating region N in the phase space of the wave. These tangencies occur when the φ component has a local minimum when (φ, u) is exterior to the rectangle M_-. The waves that we construct therefore have monotone φ components whenever (φ, u) is exterior to this region. This is in itself a rather artificial requirement, which accounts for the artificiality of the hypothesis.

It should be noted, however, that Hypotheses 1–5 provide global geometric conditions on the nonlinearities in the equations and will hold in regimes other than in the essentially decoupled regime in which λ is assumed to be small. In particular, it should be noted that all of the above hypotheses are independent of the parameter α. In section 4, we show that for large α, the u-component of the profile is monotone increasing while the φ component has a single local maximum. On the other hand, for small α, the φ-component is monotone decreasing while the u-component is either monotone increasing (if $g = 0$ is monotone) or has a single local minimum (if $g = 0$ has a local minimum). Theorem 2.1 can be viewed as a global continuation of these two asymptotic regimes, each of which is quite far from the scalar bistable travelling wave problem. Some illustrative numerical calculations are presented in section 6.

THEOREM 2.1. *Under the above hypotheses, there exists a solution of* (1) *connecting the rest point P_- at $-\infty$ to the rest point P_+ at $+\infty$ for some wave speed $c < 0$. The solution satisfies the following bounds:*

$$u_w \le u(\xi) \le u_+,$$

$$\varphi(\xi) \ge \phi_+.$$

Furthermore, if φ_m is the value of φ in the left vertical edge of M_- and if $\varphi_{**} = \min\{\varphi_m, \varphi_w\}$, then there exist unique values $\xi_m \le \xi_{**}$ such that $\varphi(\xi_m) = \varphi_m$,

$\varphi(\xi_{**}) = \varphi_{**}$, *and*

$$\dot{\varphi}(\xi) < 0 \quad (\xi \geq \xi_m),$$

$$\dot{u}(\xi) > 0 \quad (\xi \geq \xi_{**}).$$

3. Proof of Theorem 2.1. The proof employs a topological invariant called the *Conley connection index*, which is a variant of the Conley index whose formulation permits it to detect codimension-one connections, i.e., connections between rest points P_- and P_+ for which the unstable manifold of the former and the stable manifold of the latter have dimensions adding up to that of the total state space. In this situation, any intersection of such manifolds is necessarily nontransverse, and in order to detect connections topologically, it is necessary to augment the flow with a trivial parameter flow; in this case, we include the wave velocity parameter c as a new dependent variable. The "nontriviality" of the index then forces the existence of a connecting orbit for at least one value of the wave-speed parameter. A detailed discussion of the connection index can be found in Conley and Gardner [6]; see also [10].

The main analytical construction consists of finding a certain neighborhood N in the unaugmented phase space which contains both rest points in its interior and which is isolating for each c in some interval of wave speeds $c_0 \leq c \leq c_1$. Furthermore, it is required that at the extreme parameter values c_0 and c_1, there are *no* orbits connecting P_- to P_+. The construction of the neighborhood N given here closely parallels that of the example of competitive diffusion equations studied in [6] with one important difference. In [6], both components of the wave are monotone for all parameter values for which it exists. Here, however, either component may sometimes be nonmonotone, and the neighborhood N in which we expect to locate the connection must be revised accordingly. In this regard, the construction is closer in spirit to a paper on travelling wave solutions of predator–prey systems [10], where nonmonotone behavior is also encountered.

The isolating neighborhood N will be a region of the form

$$N = N_0 \cup N_- \cup N_+ \setminus N_*(\varepsilon),$$

where $N_\pm$ are neighborhoods of the rest points $P_\pm$ of the form

$$N_\pm = M_\pm \times \{\psi : |\psi| < K\},$$

N_0 is the region

$$N_0 = M_0 \times \{\psi : -K \leq \psi \leq 0\},$$

where K is a large positive constant, and $N_*(\varepsilon)$ is a small ε neighborhood of P_*. The regions $M_\pm$ and M_0 are as in Hypotheses 4 and 3, respectively. It should be noted that these regions need to vary with the parameters.

In order to show that N is isolating, we need to verify that $S(N) \cap \partial(N)$ is empty for $c_1 \leq c \leq c_0$, where $S(N)$ denotes the set of points on solutions of (1) which remain in N for all time and, furthermore, that $S(N) = \{P_-, P_+\}$ for $c = c_0, c_1$. In order to achieve this last condition, we shall choose c_1 to be a large negative constant and c_0 to be negative and close to zero.

We begin by proving a lemma which characterizes the possible solutions in $S(N)$. For convenience, we shall denote system (1) in vector form as

$$\dot{x} = f(x, c),$$

where $x = (u, \varphi, \psi)$ and c is the wave-speed parameter. Also, we shall denote the resulting flow by $x \cdot \xi$.

LEMMA 3.1. *For all $c < 0$, $S(N)$ consists of the rest-point set $P_\pm$ interior to N and orbits in N (if any) which connect P_- at $\xi = -\infty$ to P_+ at $\xi = +\infty$.*

Proof. We first show that the only orbits which remain in the neighborhoods $N_\pm$ for all time are the rest points $P_\pm$. Suppose, for example, that there is a nonconstant orbit $x \cdot \xi \in N_-$ for all ξ. Let

$$N_\tau = M_-(\tau) \times \{\psi : |\psi| < K\},$$

where $M_-(\tau)$ is as in Hypothesis 4, and set

$$\tau_0 = \inf\{\tau \in (0, 1] : x \cdot \xi \in N_-(\tau) \text{ for all } \xi\}.$$

Since $x \cdot \xi \in N_-(1)$ for all ξ and is nonconstant, it follows that $0 < \tau_0 \leq 1$. By the minimal property of τ_0, it follows that $x_1 = x \cdot \xi \in \partial N_-(\tau_0)$ for some ξ. It follows immediately from Hypothesis 4 that if (φ_1, u_1) lies in a horizontal edge of $M_-(\tau_0)$, then $\dot{u} \neq 0$ at this point so that the orbit would have to exit $N_-(\tau_0)$ in one time direction. If (φ_1, u_1) lies in a vertical edge, then if $\psi_1 \neq 0$, the solution again leaves $N_-(\tau_0)$ in one direction, while if $\psi_1 = 0$, then by Hypotheses 1 and 4, $\ddot{\varphi}(\xi)$ is positive on the right vertical edge and negative on the left vertical edge. In either case, the solution remains outside $N_-(\tau_0)$ in both time directions. Thus in all cases, the solution would exit $N_-(\tau)$ for some τ larger than but near τ_0, contradicting the defining condition for τ_0. Finally, since (φ, u) lies in M_- for all ξ, there exists L such that $|u|$ and $|\varphi|$ are uniformly bounded by L. It then easily follows from standard regularity theory for parabolic equations that there exists $K = K(L)$ such that $|\psi| < K$ for solutions for which (φ, u) lie in M_-. If K is sufficiently large, it therefore follows that $|\psi_1| < K$. Thus in all cases, we see that the only solution which can remain in N_- for all time is the constant solution P_-. The proof for N_+ is the same and will be omitted.

Now suppose that $x \in S(N)$ and that it is not a rest point. The argument above shows that

$$x \cdot \xi \in N_0 \setminus (N_- \cup N_+ \cup N_*(\varepsilon))$$

for some ξ. In the region N_0, $\varphi(\xi)$ is a monotone decreasing function. If the solution were to remain in N_0 for all ξ, it would then follow that $\varphi(\xi)$ would tend to distinct limits at $\pm\infty$. The first equation in (1) can then be viewed as an asymptotically autonomous scalar equation for u, from which it easily follows that $u(\xi)$ tends to limits at $\pm\infty$ as well. Similarly, it easily follows from the second equation in (1) that $\psi(\xi)$ tends to zero at $\pm\infty$. The only way that this can occur is if the forward and backward limits are rest points of (1); however, the only rest points of (1) in N are $P_\pm$. It therefore follows that $x \cdot \xi$ enters N_- in some backward time $\xi = \xi_-$ and that it enters N_+ in some forward time $\xi = \xi_+$.

We now claim that $x \cdot \xi$ remains in N_- (resp. N_+) for $\xi < \xi_-$ (resp. for $\xi > \xi_+$). For example, if, after having entered N_+ at ξ_+, it were to exit this set at some $\xi_* > \xi_+$, it would have to do so by having (φ, u) cross the lower edge or the right edge of M_+, i.e., an edge interior to M_0. If it were to cross the right edge, then $\psi(\xi_*) \geq 0$. If ψ were positive at this point, this would persist for $\xi > \xi_*$ as the orbit enters the region, N_0; however, in this region, ψ is nonpositive, yielding a contradiction. Also, $\psi(\xi_*)$ cannot vanish since by Hypothesis 3, we have that at such a point, $\ddot{\varphi}(\xi_*) = -F > 0$ so that $x \cdot \xi$ would be exterior to N_+ for all ξ near ξ_*, again yielding a contradiction.

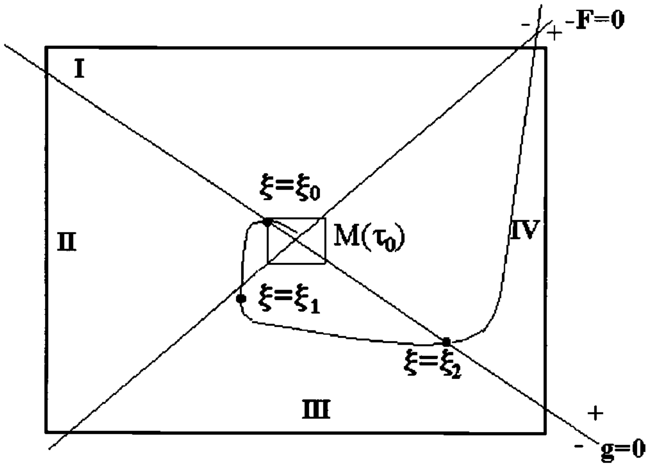


FIG. 3. *Behavior in M_+.*

If the orbit were to cross back into N_+ through the lower edge of M_+, by the above reasoning, $\dot{\varphi}$ would have to remain negative along the forward half-orbit, while $\dot{u}$ would have to remain negative (see Figure 1). This is because $g = 0$ has a negative slope here so that the monotonicity of both components must persist along the forward half-orbit. Such a solution must eventually exit N by having (φ, u) exit M_0 through the left or bottom edge.

The argument that $x \cdot \xi$ remains in N_- for $\xi \le \xi_-$ is proved by a similar, simpler argument. It will therefore be omitted.

We have now demonstrated that $\omega(x \cdot \xi) \subset N_+$ and that $\alpha(x \cdot \xi) \subset N_-$. However, the argument in the first paragraph of the proof of this lemma shows that the only way that this can occur is if $\omega(x \cdot \xi) = \{P_+\}$ and $\alpha(x \cdot \xi) = \{P_-\}$, completing the proof.

LEMMA 3.2. 1. *Suppose that $x \cdot \xi \in S(N)$ is a nonconstant orbit for system* (1). *Let u_w be the minimal value of u along $\{g = 0\} \cap M_0$; then $u_w \le u(\xi) \le u_+(\lambda)$ and $\varphi(\xi) > \varphi_+(\lambda)$ for all ξ.* 2. *Let the left vertical edge of the rectangle M_- be $\varphi = \varphi_m$, let the point where $g = 0$ has its minimum be (φ_w, u_w), and let $\varphi_{**} = \min\{\varphi_m, \varphi_w\}$. Then there exist unique points $\xi_m \le \xi_{**}$ such that $\varphi(\xi_m) = \varphi_m$ and $\varphi(\xi_{**}) = \varphi_{**}$. Furthermore, $\psi(\xi) \le 0$ for $\xi > \xi_m$ and $\dot{u}(\xi) > 0$ for $\xi > \xi_{**}$.*

Proof. 1. If $\lambda = 0$, then the solutions in $S(N)$ have constant u-component, in which case the bounds for u are obvious. Suppose then that $\lambda > 0$. The lower bound u_w for u is obvious since $\dot{u} < 0$ whenever $u < u_w$. Next, let S be the supremum of $u(\xi)$ and assume that $S > u_+$. Since $u(\xi)$ tends to limits $u_\pm < S$ at $\pm\infty$, it follows that $S = u(\xi_0)$ for some finite ξ_0 and that $\dot{u}(\xi_0) = 0$. It then follows that (φ, u) lies in the portion of the null cline $\{g = 0\}$ inside M_+ for which $u > u_+$ at $\xi = \xi_0$ (see Figure 3). We therefore also have that $\varphi(\xi_0) < \varphi_+$.

The region M_+ is composed of four wedge-shaped regions bounded by the null clines $F = 0$ and $g = 0$ and the edges of M_+, which we have denoted by I, II, III, and IV in Figure 3. At ξ_0, (φ, u) must lie on the boundary of I and II. Let τ_0 be the (unique) value of $\tau \in (0, 1)$ such that

$$(\varphi(\xi_0), u(\xi_0)) \in \partial M_+(\tau_0),$$

where $M_+(\tau)$ is the family of rectangles about P_+ defined in Hypothesis 4. Since

$S = u(\xi_0)$ is a maximum, it follows that $\ddot{u}(\xi_0) \leq 0$. However, from the first equation in (1), we see that $\ddot{u}(\xi_0) = -c\lambda W'(\varphi(\xi_0))\psi(\xi_0)$; since $-cW'(\varphi) > 0$ for $(\varphi, u) \in M_+$, it follows that $\psi(\xi_0) \leq 0$. However, if $\psi(\xi_0) = 0$, then $\dot{\psi}(\xi_0) = -F(\varphi(\xi_0), u(\xi_0)) < 0$ so that $u^{\cdots}(\xi_0) < 0$, contradicting the maximality of S for $\xi < \xi_0$. We therefore have that $\psi(\xi_0) < 0$, and the solution must therefore behave as indicated in Figure 3, i.e., (φ, u) must exit $M(\tau_0)$ in the forward time direction by moving from region I to region II. Note that both components are now decreasing monotonically with ξ.

We next claim that there exists $\xi_1 > \xi_0$ such that φ has a local minimum at ξ_1. This follows from our previous observation that $\varphi(\xi_0) < \varphi_+$ and that $x \cdot \xi \in S(N)$, so φ must have a minimum at some finite value of $\xi > \xi_0$. We assume that ξ_1 is the smallest such point. It follows that $\ddot{\varphi}(\xi_1) = -F \geq 0$ so that the point (φ, u) must lie either on the left boundary of region III or in its interior at ξ_1. In the former case, we see that $\varphi^{\cdots}(\xi_1)$ is positive since $F_u > 0$, contradicting the minimality of ξ_1. Hence the solution must lie interior to III so that $\ddot{\varphi}(\xi_1) > 0$ and thus φ has a strict local minimum at ξ_1 and $\dot{\varphi} > 0$ for $\xi > \xi_1$. Note that (φ, u) remains outside the rectangle $M_+(\tau_0)$ on the interval $\xi_0 < \xi \leq \xi_1$ and that $u(\xi_1)$ is smaller than the smallest value of u for all points in this rectangle.

For $\xi > \xi_1$ but not too large, it follows that u must continue to decrease and φ must continue to increase with ξ. Clearly, this monotonicity cannot persist for all ξ since $u(\xi_1) < u_+$, and by the previous lemma, the solution must tend to P_+ at $+\infty$. The only way that this can occur is for (φ, u) to cross $g = 0$ at a time $\xi = \xi_2$ by crossing from region III into region IV. Furthermore, φ must remain monotone increasing since (φ, u) remains below $F = 0$ on this interval.

Finally, we have that at some $\xi = \xi_2 > \xi_1$, (φ, u) lies on $g = 0$, defining the common boundary of III and IV; furthermore, it is still exterior to the rectangle $M_+(\tau_0)$. We now observe that both φ and u must continue to increase while the solution remains in region IV since $\dot{u} > 0$ and $\ddot{\varphi} > 0$ in this region. Since the orbit ultimately tends to P_+, this monotonicity cannot persist for all $\xi \geq \xi_2$, and so the orbit must cross into region I from region IV through their common boundary at some time $\xi_3 > \xi_2$. However, by monotonicity, we again have that (φ, u) remains exterior to $M(\tau_0)$ for $\xi_2 \leq \xi \leq \xi_3$. Thus at $\xi = \xi_3$, u attains a value larger than $S = u(\xi_0)$, yielding a contradiction.

It is also easily proved that $\varphi(\xi) > \varphi_+$ for all ξ. Let m be the infimum of $\varphi(\xi)$ and suppose that $m < \varphi_+$. There exists (finite) ξ_0 such that $m = \varphi(\xi_0)$, and at this point, ψ must vanish. It easily follows that (φ, u) lies in region III of M_+ so that $\dot{u}(\xi_0) < 0$ and φ has a strict local minimum here. Since the solution is assumed to lie in $S(N)$, arguments similar to those in the preceding paragraphs show that (φ, u) must eventually cross from region III to region IV and finally from region IV to region I in the forward time direction, contradicting the upper bound $u \leq u_+$ obtained above.

2. Since the orbit is a connection from P_- to P_+, there exists some $\xi = \xi_m$ such that $\varphi(\xi_m) = \varphi_m$; we assume that ξ_m is the largest value of ξ where this occurs so that $\varphi(\xi) < \varphi_m$ for $\xi > \xi_m$. It follows that $\psi(\xi_m) \leq 0$. If $\psi(\xi_m) = 0$, then $\dot{\psi}(\xi_m) = -F < 0$ since $\varphi = \varphi_m$ here. However, we would then have that φ has a strict local maximum at ξ_m so that $\psi(\xi) > 0$ and $\varphi(\xi) < \varphi_m$ for $\xi < \xi_m$. Such solutions exit N in the backward time direction. Thus $\psi(\xi_m) < 0$. If $\varphi(\xi_1) = \varphi_m$ for some $\xi_1 < \xi_m$, then assuming that ξ_1 is the largest such value less than ξ_m, we would have that $\psi(\xi_1)$ that either $\psi(\xi_1) > 0$ or $\psi(\xi_1) = 0$. Both conditions imply that the solution leaves N in the backwards direction so that ξ_m is unique.

For $\xi > \xi_m$, the solution enters the region $N \setminus (N_- \cup N_+)$ and thus $\psi(\xi) \leq 0$

for $\xi \geq \xi_m$ up to the first time $\xi = \xi_0 > \xi_m$ that the solution enters N_+, and the only possibilities are that (φ, u) enters M_+ through either region III or IV. However, if it crosses into region III at some $\xi \geq \xi_0$, then $u < u_+$, and the only way that u can begin to increase is for φ to have a local minimum at some $\xi_1 > \xi_0$ so that the solution crosses from region III into region IV for some $\xi_2 > \xi_1$. However, φ must continue to increase for as long as the solution remains in region IV, so as in part 1 of the proof, u must eventually attain values that exceed u_+—a contradiction. The only possibility is that (φ, u) enters region IV with $\dot{u} > 0$ and $\dot{\varphi} < 0$. If this monotonicity does not persist for all $\xi > \xi_0$, then φ must have a strict local minimum at some $\xi_1 > \xi_0$. For $\xi > \xi_1$, either the solution exits M_+ with $\psi > 0$ so that $x \cdot \xi$ leaves $S(N)$ or the solution remains in M_+, in which case it must cross from region IV into region I. In either case, we obtain a contradiction. We have therefore shown that ψ remains nonpositive along the entire forward half-orbit, $\xi \geq \xi_m$.

It follows from the nonpositivity of ψ for $\xi \geq \xi_m$ that $\varphi(\xi)$ is strictly monotone decreasing on this interval since if this were not the case, $\psi(\xi)$ would necessarily vanish identically on some interval. It would then follow that the orbit in question is a rest point, contrary to our assumption. It therefore follows that there exists a unique value $\xi = \xi_{**} \geq \xi_m$ such that $\varphi(\xi_{**}) = \varphi_{**}$, where $\varphi_{**} = \min\{\varphi_m, \varphi_w\}$.

We claim that $\dot{u}(\xi) > 0$ for all $\xi > \xi_{**}$. Since $\dot{\varphi} \leq 0$ for $\xi \leq \xi_{**}$, it follows that (φ, u) lies above $g = 0$ at ξ_{**} since otherwise (φ, u) would enter and remain in the region below $g = 0$ for $\xi > \xi_{**}$, and u would therefore have to decrease for all larger ξ since $g = 0$ is a monotone decreasing function of φ here; this is a contradiction. If the claim that $\dot{u} > 0$ for all $\xi \geq \xi_{**}$ were false, the point (φ, u) would have to cross the null set $g = 0$ at some $\xi_1 > \xi_{**}$. Since $\psi \leq 0$ for $\xi > \xi_m$, the φ component must decrease monotonically for all such ξ. The tangent vector to (φ, u) must therefore point into the left half-plane. Furthermore, we have that $\dot{u}(\xi_1) = 0$ so that the tangent to (φ, u) is the vector $(\psi(\xi_1), 0)$ with $\psi(\xi_1) \leq 0$. If ψ is strictly negative at this point, the solution enters the region below $g = 0$. Since φ is monotone decreasing for all $\xi \geq \xi_1$ and $g = 0$ is a monotone decreasing function of φ for $\varphi \leq \varphi_w$, it follows that (φ, u) can never cross $g = 0$ in the forward direction, and u would therefore have to monotonically decrease to a limit strictly less than u_+ at $+\infty$—a contradiction. If $\psi(\xi_1) = 0$, then we must have that $F(\varphi, u) \neq 0$ at ξ_1; otherwise, $x \cdot \xi_1$ would coincide with the rest point P_*, which is impossible. Since the solution is exterior to $N_\pm$ at this point, we must have that ψ is nonpositive near ξ_1 so that the only remaining alternative is that $-F < 0$ at this point. However, the argument now proceeds as when $\psi(\xi_1) < 0$, and the solution must still eventually leave N in the forward direction. Hence we must have that $\dot{u} > 0$ for all $\xi > \xi_{**}$.

Lemma 3.3. *There exists (large) $K > 0$ and (small) $\varepsilon > 0$ such that $S(N) \cap \partial(N)$ is empty for all $c < 0$.*

Proof. Suppose that $x \in S(N) \cap \partial N$. The boundary of N consists of the union of the following subsets:

B_1: $(\varphi, u) \in \partial M_0$ and $\psi \leq 0$;
B_2: $\psi = -K$ and $(\varphi, u) \in M_0$, or $\psi = +K$ and $(\varphi, u) \in M_\pm$;
B_3: $\psi \geq 0$ and $(\varphi, u) \in \partial M_\pm$;
B_4: $\psi = 0$ and $(\varphi, u) \in M_0 \setminus (M_+ \cup M_-)$;
B_5: $x \in \partial N_*(\varepsilon)$.

1. Suppose that $x \in B_1$. If (φ, u) lies in either horizontal edge of M_0, then $\dot{u} \neq 0$ so that the solution immediately exits N in one time direction. Next, suppose that (φ, u) lies in a vertical edge of M_0. If $\psi < 0$, then the solution immediately exits N in one

time direction. Suppose then that $\psi = 0$ at this point so that $\ddot{\varphi} = \dot{\psi} = -F(\varphi, u)$. If (φ, u) lies in the left (resp. right) edge of M_0, then $-F(\varphi, u) < 0$ (resp. $-F(\varphi, u) > 0$) so that φ has a local maximum (resp. minimum). Thus in both cases, the solution immediately exits N in both time directions.

2. Suppose that $x \in B_2$ so that $|\psi| = K$. If $x \in S(N)$ and if

$$\Phi(x,t) = \varphi(x - ct), \qquad U(x,t) = u(x - ct),$$

then Φ is a solution of the scalar parabolic equation

$$\Phi_t = \Phi_{xx} + F(\Phi, U).$$

Since the last term is a uniformly bounded, smooth function of (x, t), it follows from the $(1 + \delta)$ Schauder estimates for scalar, linear parabolic equations that Φ_x is uniformly bounded on finite time intervals so that $\Phi_x(x, 1)$ is a uniformly bounded function of x. However, $\dot{\varphi}(x) = \Phi_x(x + c, 1)$ so that $|\psi|$ is uniformly bounded from above by some constant K which depends only on the uniform bound for $F(\varphi, u)$ for $(\varphi, u) \in M_0$. In particular, K is independent of the wave speed c. It follows that for such K, if $|\psi| = K$, then $x \notin S(N)$.

3. Suppose that $x \in B_3$. If (φ, u) lies in the horizontal edges of M_-, then by Hypothesis 3, $\dot{u} \neq 0$ so that the solution immediately exits N in one time direction. If (φ, u) lies on the right vertical edge of M_- and $\psi > 0$, then (φ, u) exits M_0, and hence N, in forward time. If $\psi > 0$ with (φ, u) in the left vertical edge of M_-, then (φ, u) enters $M_0 \setminus (M_- \cup M_+)$ in forward time with $\psi > 0$; such solutions also leave N. Finally, if (φ, u) lies in a vertical edge of M_- with $\psi = 0$, then by Hypothesis 3, the solution exits N in both time directions.

Next, suppose that $(\varphi, u) \in \partial M_+$ with $\psi \geq 0$. The argument that the solution leaves N in at least one time direction is the same as in the case for $(\varphi, u) \in M_-$, with the exception of the points (φ, u) in the lower horizontal edge of M_+ when $\psi = 0$ since the latter edge is interior to M_0. However, at such a point, we have that $\dot{u} < 0$ with $\varphi < \varphi_{**}$; by Lemma 3.2, the solution through such a point cannot lie in $S(N)$.

4. Suppose that $x \in B_4$ so that $\psi = 0$ and $(\varphi, u) \in M_0 \setminus (M_- \cup M_+)$. If $F(\varphi, u) \neq 0$, then $\dot{\psi} = -F(\varphi, u)$ is nonvanishing so that ψ becomes positive in one time direction while $(\varphi, u) \notin (M_- \cup M_+)$; such solutions cannot lie in $S(N)$. We therefore have that $F = 0$ at such points. Since $(\varphi, u) \notin M_\pm$, it follows that (φ, u) must lie in either the middle branch of $F = 0$ or the portion of the left branch outside M_+. However, in the latter case, $\varphi < \varphi_+$, which is impossible by Lemma 3.2. Hence (φ, u) must lie in the middle branch of $F = 0$. If (φ, u) lies strictly below $g = 0$ at x, then $\dot{u} < 0$ and $\ddot{\psi} = -F_u \dot{u} > 0$ so that $\psi > 0$ for small ξ and the solution leaves N in both directions. Furthermore, if $g(\varphi, u) = 0$, then $x = P_*$ which is exterior to N—also a contradiction. Hence we must have that (φ, u) lies strictly above $g = 0$ at x.

By part 2 of Lemma 3.2, $\psi(\xi)$ remains nonpositive along the backward orbit segment $x \cdot [\xi_m, 0]$, beyond which point the orbit enters N_-. Furthermore, by an argument in the previous paragraph, $\psi < 0$ on this interval whenever $F(\varphi, u) \neq 0$. If $\psi = 0$ at some $\xi < \xi_m$, let ξ_* be the largest value of $\xi \leq \xi_m$ for which $\psi = 0$; if $\psi < 0$ for all $\xi < \xi_m$, let $\xi_* = -\infty$. Let $\varphi_r^+ = h_r(u_+(\lambda))$. We claim that $\varphi(\xi_*) \in [\varphi_-, \varphi_r^+]$. The upper bound follows immediately from the estimate $u(\xi) \leq u_+(\lambda)$ of Lemma 3.2. We next obtain the lower bound for $\varphi(\xi_*)$. In the event that $\xi_* = -\infty$, we have that $\varphi(\xi_*) = \varphi_-$. Assume then that ξ_* is finite and that $\varphi(\xi_*) < \varphi_-$. A contradiction will be obtained in each of two separate cases, (a) and (b), below.

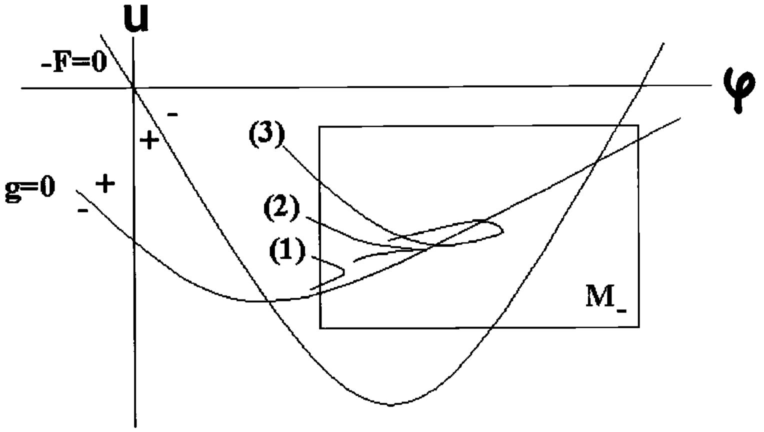


FIG. 4. *Possible behavior of the backwards orbit with* $\psi(0) = 0$.

(a) Suppose that the graph of $g = 0$ is a monotone decreasing function of φ in the range $\varphi_+ \le \varphi \le \varphi_-$. We claim that (φ, u) lies strictly above $g = 0$ on the interval $[\xi_*, 0]$. As noted above, this condition holds at $\xi = 0$. If the claim is false, let $\xi_1 \in [\xi_*, 0]$ be the largest ξ in this interval for which $g(\varphi, u) = 0$. If $\psi = 0$ at ξ_1, then by the conditions above, we must have that $F = 0$ as well, in which case x must be a rest point, and the only possibility is that $x = P_*$. However, P_* is exterior to N. Hence we must have that $\psi < 0$ at ξ_1, in which case the tangent vector to (φ, u) is $(\psi(\xi_1), 0)$. Since we have assumed that $\varphi < \varphi_-$ at ξ_*, it follows by monotonicity that $\varphi(\xi_1) < \varphi_-$ as well, and since $g = 0$ is a monotone decreasing curve, this contradicts the maximality of ξ_1. Thus $g(\varphi, u) > 0$ on $[\xi_*, 0]$. If $\varphi(\xi_*) < \varphi_-$, then by the monotonicity of $g = 0$, it follows that $\varphi(\xi_*)$ is a strict local maximum so that the solution would then necessarily exit N in the backwards direction by leaving N_- while $\psi > 0$. It therefore follows that that $\varphi(\xi_*) > \varphi_-$.

(b) Next, suppose that $g = 0$ has its local minimum at the point (φ_w, u_w) with $\varphi_w \in (\varphi_+, \varphi_-)$. By the definition of ξ_*, we have that $\varphi(\xi_*) > \varphi_m$. There are three distinct possibilities, as depicted in Figure 4. In case 1, the point (φ, u) lies above $g = 0$ at ξ_*. Since $\ddot{\varphi} = -F < 0$ at this point, φ must have a local maximum at this point and $\psi > 0$ for $\xi < \xi_*$. It now follows that the orbit cannot cross the monotone increasing portion of $g = 0$ for $\xi < \xi_*$. It follows that (φ, u) must remain in the region $-F < 0$ for as long as the backward orbit remains in N_-, and it follows that in this region, ψ must remain strictly positive. Thus φ decreases for $\xi < \xi_*$ until the solution leaves N_-. At this point, ψ is still positive, and the solution therefore exits N in the backward direction.

The other two possibilities are depicted by the orbits labelled 2 and 3 in Figure 4. In the former case, (φ, u) lies on $g = 0$ at ξ_* and the curve has a cusp at this point. However, the orbit still exits in the backward time direction. The argument is the same as for orbit 1. In the last case, case 3, (φ, u) lies below $g = 0$ at ξ_*. In this case, the positivity of ψ is again preserved for $\xi < \xi_*$ and the curve (φ, u) must cross $g = 0$ again before leaving N_-. As before, the orbit is seen to leave N in the backward direction since the orbit leaves N_- while $\psi > 0$.

We have now established the claim that $\varphi(\xi_*) \in [\varphi_-, \varphi_r^+]$. Since $\hat{\varphi}_- \le \varphi_-$, we then have that

$$\hat{\varphi}_- \leq \varphi(\xi_*) \leq \varphi_r^+,$$

where $\hat{\varphi}_- = h_r(u_w)$ and $\varphi_r^+ = h_r(u^+)$ are as in Hypothesis 5. Also, since $\varphi(0)$ lies on the middle branch of $F = 0$ and $u(0) < u_+(\lambda)$, it follows that

$$\varphi_*^+ \leq \varphi(0) \leq \hat{\varphi}_*$$

so that by Hypothesis 5, $R(\varphi(0)) > R(\varphi(\xi_*))$ (see also Figure 2). By Lemma 3.2, we have that $u(\xi) \geq u_w$, and by Hypothesis 1, we have that $F_u > 0$; it therefore follows that

$$\begin{aligned} \ddot{\varphi} &= -\alpha c\dot{\varphi} - F(\varphi, u) \\ &\leq -\alpha c\dot{\varphi} - F(\varphi, u_w). \end{aligned}$$

Since $\dot{\varphi} \leq 0$ on $[\xi_*, 0]$ and vanishes at the endpoints, we may multiply the above by $\dot{\varphi}$ on this interval and integrate to obtain

$$\begin{aligned} 0 &= \int_{\xi_*}^{0} (\dot{\varphi}(\xi)^2/2)^{\cdot} \, d\xi \\ &\geq -c\alpha \int_{\xi_*}^{0} \dot{\varphi}^2 \, d\xi + R(\varphi(0)) - R(\varphi(\xi_*)), \end{aligned}$$

where $R(\varphi)$ is an antiderivative of $-F(\varphi, u_w)$. We have by Hypothesis 5 and the negativity of c that both quantities in the last term of the above inequality are positive, yielding a contradiction. Thus $B_4 \cap S(N)$ is empty.

5. Finally, suppose that $x_\varepsilon \in B_5$ lies in $S(N)$ for a sequence ε tending to zero. Since each such orbit is a connection from P_- to P_+, there exists $\xi_\varepsilon > 0$ such that $(\varphi, u)(-\xi_\varepsilon)$ lies in the left vertical edge of M_- and is hence uniformly bounded away from the rest points. Let $y_\varepsilon = x_\varepsilon \cdot (-\xi_\varepsilon)$; by passing to a suitable subsequence, it can be assumed that y_ε converges to a limit $\bar{y}$. Furthermore, since x_ε converges to the rest point at P_*, it follows that $\xi_\varepsilon \to +\infty$. Since $\psi < 0$ whenever $(\varphi, u) \in M_0 \setminus (M_- \cup M_+)$, it follows that the orbit $\bar{y} \cdot \xi$ must remain in the region $\varphi \geq \varphi_*$ for all positive ξ, and since φ is monotone along the forward half-orbit, the solution must be a connection between P_- and P_*. It also follows, as in case 4, that the orbit must lie in the region above $g = 0$ as long as (φ, u) is exterior to M_-. Set $\xi_* = -\infty$ if $\psi < 0$ along the entire orbit; otherwise, set ξ_* to be the maximal $\xi < +\infty$ where ψ vanishes. As in case 4, it follows that $\varphi(\xi_*) \in [\varphi_-, \varphi_r^+]$. The proof is the same as the previous case and it will therefore be omitted.

A contradiction is obtained as before by replacing u with u_w in the $\dot{\psi}$ equation, multiplying by $\dot{\psi}$, and integrating over the interval $[\xi_*, +\infty)$. This completes the proof of the lemma.

We now have an isolating region N for negative wave speeds c which contains precisely two rest points, P_- and $P_+(\lambda)$. Furthermore, the only nonconstant solutions in N are connections between these two rest points. In order to complete the proof of Theorem 2.1, we need to use N to construct connection triples for the system augmented with the parameter flow $\dot{c} = 0$. In particular, an interval $I = [c_1, c_0]$ must be determined such that $S(N) = \{P_-, P_+(\lambda)\}$ when $c = c_0, c_1$. The interval I must be large enough to contain the wave speed $c < 0$ for which the connecting orbit occurs;

in this case, that means choosing $|c_1|$ sufficiently large and $c_0 < 0$ sufficiently close to zero.

LEMMA 3.4. (a) *There exits $c_0 < 0$ such that $S(N) = \{P_-, P_+(\lambda)\}$ for all $c \in [c_0, 0)$.* (b) *There exists $c_1 \ll 0$ such that $S(N) = \{P_-, P_+(\lambda)\}$ for all $c < c_1$.*

Proof. (a) When $c = 0$, u is constant and negative. Let $U = \max\{u_+(\lambda) : 0 \leq \lambda \leq \lambda_0\}$ so that $u \in [u_w, U]$ for $0 \leq \lambda \leq \lambda_0$. For u in this range, the (φ, ψ) equations form a one-parameter family of Hamiltonian "fisheye" systems which are all topologically equivalent, which have an orbit homoclinic to the rest point P_-, and which encircle P_*. In particular, for such a u, there is never an orbit running from a rest point on the right branch, $\varphi = h_r(u)$, to the corresponding rest point on the left branch, $\varphi = h_\ell(u)$, of $F = 0$. Let V be a small tubular neighborhood of the two curves of rest points,

$$(u, \varphi, \psi) = (u, h_{r,\ell}(u), 0),$$

of the $c = 0$ system. It then follows that at $c = 0$, there exists $T > 0$ such that if $x \in N \setminus V$, then at least one of $x \cdot (\pm T)$ is not in N. It then follows from standard continuous-dependence theorems for flows that at least one of $x \cdot (\pm T)$ is not in N for all sufficiently small $c < 0$ and $x \in N \setminus V$. However, any connecting orbit from P_- to P_+ necessarily enters the region $N \setminus V$ at some point. Hence there are no P_--to-P_+ connections for sufficiently small c.

(b) Let $\delta = c^{-1}$ and

$$U(\varphi) = u_- + \lambda(W(\varphi) - W(\varphi_-));$$

since N is a compact region, it follows that for any solution (u, φ, ψ) in $S(N)$, we have that $u(\xi) = U(\varphi(\xi)) + O(\delta)$ for sufficiently small δ. It therefore follows that such solutions have (φ, ψ) components which satisfy the system

$$\begin{aligned} \dot{\varphi} &= \psi, \\ \dot{\psi} &= -\alpha c \psi - F(\varphi, U(\varphi)) + O(\delta). \end{aligned}$$

Next, let $\gamma = \dot{\psi}$; then (γ, φ) satisfies the system

$$\begin{aligned} \delta\dot{\gamma} &= -\alpha\gamma + O(\delta), \\ \dot{\varphi} &= -\delta(F(\varphi, U(\varphi)) + O(\delta) + \gamma)/\alpha. \end{aligned}$$

It follows that $|\gamma|$ must remain $O(\delta)$ along bounded solutions so that after rescaling time, the φ component must satisfy the scalar equation

$$\dot{\varphi} = F(\varphi, U(\varphi)) + O(\delta).$$

Now the function $F(\varphi, U(\varphi))$ is qualitatively a cubic with attracting rest points at $\varphi_\pm$ and a repelling rest point at φ_*. It is therefore possible to find positively invariant neighborhoods of the first two rest points for all sufficiently small δ. It follows that there are no P_--to-P_+ connections for sufficiently large $|c|$. This completes the proof of the lemma.

The proof of Theorem 2.1 is now easily completed. Since N is isolating for all c in the interval $I = [c_1, c_0]$, it follows that a connection triple (S_-, S_+, S) is determined by the augmented flow

$$\begin{aligned} \dot{x} &= f(x, c), \\ \dot{c} &= 0, \end{aligned}$$

where $S_{\pm} = \{P_{\pm}\}\times I$, and $S = S(N\times I)$ for the augmented flow (see [6]). Furthermore, by our hypotheses, these connection triples are all related by continuation for $\lambda \in [0, \lambda_0]$ so that the connection index $\bar{h}$ of the triple is independent of λ. At $\lambda = 0$, the equations nearly decouple since the u equation

$$\dot{u} = -c(u - u_-)$$

is independent of φ. The above equation is linear with a repelling rest point at $u = u_-$ so that $u \equiv u_-$ is its only bounded solution. Thus at $\lambda = 0$, $u = u_-$ along solutions in S. Next, replace u in the argument of F in the $\dot{\psi}$ equation by

$$\sigma u_- + (1 - \sigma)u.$$

By the previous remark, the invariant set isolated by $N \times I$ is independent of $\sigma \in [0, 1]$ so that the triples isolated by $N \times I$ are all related by continuation for such σ. At $\sigma = 1$, the equations completely decouple—

$$\begin{aligned}
\dot{u} &= -c(u - u_-),\\
\dot{\varphi} &= \psi,\\
\dot{\psi} &= -c\alpha\psi - F(\varphi, u_-),
\end{aligned}$$

—and by the product formula for the connection index (see [6]), we have that $\bar{h}$ is the smash product of the connection index for the (φ, ψ) system with a 1-sphere. Since the index of the latter system is the homotopy type $[\bar{0}]$ of a point (see the appendix in [6]), it follows that $\bar{h} = [\bar{0}] \wedge \Sigma^1 = [\bar{0}]$. It therefore follows that $S(N)$ contains more than the two rest points P_- and P_+ for some $c \in I$, and by Lemma 3.2, the nonconstant solution in N must be the desired connecting orbit.

4. Singular limits I: Small and large $\boldsymbol{\alpha}$. In this section, we examine two asymptotic regimes wherein the parameter α is either small or large. For both regimes, we obtain two classes of results: a local theory in the neighborhood of certain reduced problems, wherein the existence of a wave and its uniqueness relative to some neighborhood of the reduced wave are obtained, and a global result concerning the asymptotic behavior as $\alpha \to 0$ and uniqueness relative to all possible wave speeds and all possible wave solutions in the isolating region N constructed in the previous section. In each case, the local analysis, which is based upon geometric singular perturbation theory, holds under substantially weakened hypotheses, wherein only Hypotheses 1 and 2 are required. The global theory draws upon several results in the preceding section, and therefore all five hypotheses need to be satisfied in this case.

Let $\theta = \alpha c$; system (1) can then be written as

$$\begin{aligned}
\dot{u} &= -cg(\varphi, u),\\
\dot{\varphi} &= \psi,\\
\dot{\psi} &= -\theta\psi - F(\varphi, u).
\end{aligned} \tag{2}$$

The system therefore has two independent parameters c and θ related to the wave speed. The following lemma obtains a region in parameter space for which connecting solutions can exist. This is needed to establish the global uniqueness of the wave and the wave velocity since invariant-manifold methods only provide local information about uniqueness in a neighborhood of the singular limit. We remark that if we do not impose some particular ansatz for the asymptotic limit, c is an arbitrary negative parameter so that even if α is assumed to be small, θ is also an arbitrary negative parameter.

LEMMA 4.1. *There exists $\theta_0 < 0$ such that there are no solutions of* (2) *in the isolating region N which connect P_- to P_+ for* any *$c \leq 0$ and $\theta \in [\theta_0, 0]$.*

Proof. For large c, we expect that $g = 0$ along bounded solutions of (2). Hence we set

$$U(\varphi) = u_- - \lambda(W(\varphi) - W(\varphi_-)),$$
$$h(\varphi) = F(\varphi, U(\varphi))$$

and consider the reduced problem

$$\begin{aligned} \dot{\varphi} &= \psi, \\ \dot{\psi} &= -\theta\psi - h(\varphi), \\ u(\xi) &= U(\varphi(\xi)). \end{aligned} \tag{3}$$

It follows from Hypotheses 1 and 2 that $h(\varphi)$ has three distinct roots $p_+ < p_* < p_-$. Furthermore, if $H'(\varphi) = h(\varphi)$, then it follows from Hypothesis 1 and, in particular, from $F_u < 0 \in M_0$ that $H(p_-) < H(p_+)$. In this manner, we see that (3) has a unique connection from p_- to p_+ which occurs for a unique negative value of $\theta = \theta_R < 0$. It is well known that this connection occurs as the transverse intersection of the center-unstable manifold of $(p_-, 0, 0)$ with the center-stable manifold of $(p_+, 0, 0)$ after appending the parameter flow $\dot{\theta} = 0$ to (3).

Next, rescale ξ to $y = |c|\xi$ and set $\delta = 1/|c|$; system (3) then takes the form

$$\begin{aligned} u' &= g(\varphi, u), \\ \varphi' &= \delta\psi, \\ \psi' &= \delta(-\theta\psi - F(\varphi, u)), \\ \theta' &= 0. \end{aligned} \tag{4}$$

Let $\mathcal{M}_0 = \{(U(\varphi), \varphi, \psi, \theta)\}$ so that $\mathcal{M}_0$ is a manifold of rest points of (4) when $\delta = 0$. Furthermore, since $g_u = 1$, $\mathcal{M}_0$ is normally hyperbolic in the sense of Fenichel [7]. It follows that $\mathcal{M}_0$ perturbs smoothly for small δ to an invariant manifold $\mathcal{M}_\delta$, where the flow on the perturbed manifold is given by (3) with $U(\varphi)$ replaced by $U(\varphi) + O(\delta)$. It immediately follows that a connection exists for small delta and that the wave speed $\theta = \theta_R + O(\delta)$. Furthermore, given any $\eta > 0$, let V_η be an η neighborhood (in $\mathbf{R}^4$) of $\mathcal{M}_0$. Since the p_--to-p_+ connection occurs as a transverse intersection of invariant manifolds of the augmented equations at $\delta = 0$, it follows that for sufficiently small η, there exists $\delta_1 > 0$ such that the connection lies in V_η for $0 \leq \delta \leq \delta_1$ and that the connecting solution is unique relative to this neighborhood.

Let $c_1 = -1/\delta_1$ and let $\theta \in (\theta_1, 0)$, where $\theta_1 = \theta_R/2$. Also, set η as in the previous paragraph. Now suppose that we have a connecting solution of (2) in the isolating neighborhood N of Theorem 2.1 for some $c \leq c_1$ and $\theta \in [\theta_1, 0]$. Changing to the scaling in (4), we find that $(u - U)' = g(\varphi, u) + O(\delta)$ so that when $u = U \pm \eta$, $(u - U)'$ has the same sign as $g(\varphi, U(\varphi) \pm \eta)$. It follows from this that V_η is a positively invariant set for (4) for sufficiently small δ so that if $|u - U(\varphi)| \geq \eta$ at some point along the solution, then this condition persists for all ξ in the forward time direction. It follows that u must be unbounded along such a solution, leading to a contradiction. Thus we must have that *all* connecting solutions relative to N lie interior to V_η for such c. Hence the only possible connecting solution in N is the one lying near the invariant manifold $\mathcal{M}_0$. However, for c in the specified range, the wave speed θ of such a connection must approximate $\theta_R = 2\theta_1$ to within $O(\delta)$; hence the set of connections relative to N is empty for $c \le c_1$ and $\theta \in [\theta_1, 0]$.

Next, set $c = 0$ in (2) so that (φ, ψ) solve the last two equations with u constant. By Lemma 3.2, solutions in $S(N)$ satisfy $u \in [u_w, u_+] = I_u$. For each fixed u in this range, (2) has three distinct rest points $p_\pm(u)$ and $p_*(u)$, and there is a connecting solution from $p_-(u)$ to $p_+(u)$ which exists at some unique $\theta = \theta(u)$, where the continuous function $\theta(u)$ is strictly negative on I_u. Let $\theta_2 < 0$ be greater than the minimum of $\theta(u)$ over I_u so that $\theta_2 < 0$. The phase plane of (2) at $c = 0$ for each $u \in I_u$ consists of the two saddles at $p_\pm$ and either an unstable node or a spiral at p_*. For $\theta \le \theta_2$, there are no connecting orbits from p_- to p_+. For such a θ, it immediately follows that one of the following hold for every nonconstant solution in N: φ becomes unbounded in at least one time direction or ψ becomes positive at some point where $\varphi < \varphi_m$, where φ_m is as in Lemma 3.2. In either case, such solutions exit N in finite time. It follows from regular perturbation theory that there exists $c_2 < 0$ such that $S(N) = \{P_\pm\}$ for $c \in [c_2, 0]$ and $\theta \in [\theta_2, 0]$.

If $c_2 < c_1$ we are done. Suppose then that $c_1 < c_2$. Finally, set $\theta = 0$ in (2) and suppose that $c \in [c_1, c_2]$. It follows from Lemma 3.1 that nonconstant solutions in $S(N)$ are connections from P_- to P_+. Furthermore, by Lemma 3.2, we have that $u(\xi) \le u_+$ along any such solution, and by Hypothesis 1, we have that $F(\varphi(\xi), u(\xi)) \le F(\varphi(\xi), u_+)$. Also, by Lemma 3.2, we have that $\psi < 0$ on a (maximal) interval of the form $(\xi_0, +\infty)$, where $\xi_0 \ge -\infty$. Thus if ξ_0 is finite, then $\psi = 0$ at this point. It follows from the above for $\xi \in (\xi_0, +\infty)$ that

$$\dot{\varphi}\ddot{\varphi} \le -F(\varphi, u_+)\dot{\varphi}.$$

Let H(φ) be an antiderivative of $-F(\varphi, u_+)$; by Hypothesis 1, we have that $H(\varphi_+) < H(\varphi)$ for all $\varphi > \varphi_+$ since $u_+ < 0$. Integration of this inequality on the interval $(\xi_0, +\infty)$ yields the inequality

$$0 \le H(\varphi_+) - H(\varphi(\xi_0)),$$

and by the conditions above, the quantity on the right is negative, yielding a contradiction. Thus there are no connecting orbits when $\theta = 0$ and $c \in [c_1, c_2]$. It immediately follows from regular peturbation theory that there exists $\theta_3 < 0$ such that the set of connections is empty for $\theta \in [\theta_3, 0]$ and $c \in [c_1, c_2]$.

The proof is completed by taking θ_0 to be the (negative) maximum of θ_i, $i = 1, 2, 3$.

We can now prove the main theorems addressing the asymptotic behavior of the connecting solutions of (1) for small and large α. We first consider the case of small α and, accordingly, write (4) in the equivalent form,

$$\begin{aligned} u' &= -\theta g(\varphi, u), \\ \varphi' &= \alpha\psi, \\ \psi' &= \alpha(-\theta\psi - F(\varphi, u)), \\ \theta' &= 0, \end{aligned} \tag{5}$$

where "prime" is $\frac{d}{dy}$ with $y = \alpha^{-1}\xi$.

THEOREM 4.2. (a) *Suppose that Hypotheses* 1 *and* 2 *are satisfied. There exists* $\alpha_0 > 0$ *such that for* $\alpha \in (0, \alpha_0)$ *and* $c = \alpha\theta$*, there exists a solution* $(u(y, \alpha)$, $\varphi(y, \alpha)$,

$\psi(y,\alpha)$, $\theta(\alpha)$) of (5) *which is near the connecting solution $(\varphi_R(\xi),\psi_R(\xi),\theta_R)$ of the reduced equation* (3) *in the sense that the perturbed solution satisfies the estimates*

$$\begin{aligned} u(\alpha^{-1}\xi,\alpha) &= U(\varphi_R(\xi)) + O(\alpha), \\ \varphi(\alpha^{-1}\xi,\alpha) &= \varphi_R(\xi) + O(\alpha), \\ \psi(\alpha^{-1}\xi,\alpha) &= \psi_R(\xi) + O(\alpha), \\ \theta(\alpha) &= \theta_R + O(\alpha) \end{aligned}$$

modulo a phase shift in ξ. The perturbed solution is unique relative to some neighborhood of the reduced solution (in its four-dimensional phase space), and it depends smoothly on parameters in the equations.

(b) *Suppose that all five hypotheses in section* 2 *are satisfied. Then the perturbed solution obtained in* (a) *is unique relative to all possible solutions in the isolating region N constructed in section* 3 *and all negative wave speeds $c < 0$.*

Proof. (a) Clearly, the parameter α in (5) plays the same role as the parameter δ in (4). An invariant manifold $\mathcal{M}_\alpha$ of the perturbed equations can be constructed near the invariant manifold $\mathcal{M}_0$ of (3) as in the proof of Lemma 4.1, and the proof of part (a) follows as in the previous lemma by transversality.

It was also proved in Lemma 4.1 that in this regime, all connecting solutions in N of (4) lie in the η-neighborhood V_η of $\mathcal{M}_0$. Furthermore, by the same lemma, it follows that such connections exist only for $\theta < \theta_0 < 0$, which is in the range of parameter values for which the manifold $\mathcal{M}_0$ will be normally hyperbolic for (5) with $\alpha = 0$. Hence any such connection must lie in the manifold $\mathcal{M}_\alpha$ for sufficiently small α, and it must therefore coincide with the unique solution in this invariant manifold.

The second result concerns asymptotic behavior as α tends to $+\infty$. We shall see that the wave is resolved into an inner and an outer layer, which makes the analysis more complicated than the previous limit $\alpha = 0$, where the wave is contained entirely in the slow manifold $\mathcal{M}_\alpha$. Set $1/\alpha = \delta$ so that (1) can be written as

$$\begin{aligned} \dot{u} &= -\delta\theta g(\varphi,u), \\ \dot{\varphi} &= \psi, \\ \dot{\psi} &= -\theta\psi - F(\varphi,u) \end{aligned} \tag{6}$$

and also in rescaled form as

$$\begin{aligned} u' &= -\theta g(\varphi,u), \\ \delta\varphi' &= \psi, \\ \delta\psi' &= -\theta\psi - F(\varphi,u), \end{aligned} \tag{7}$$

where $' = \frac{d}{dy}$ and $y = \delta\xi$. As $\delta \to 0$, a matched asymptotic expansion of the solution is constructed in the usual way. The transition layer is the solution $(u_+, \varphi_F(\xi), \psi_F(\xi), \theta_F)$ of (7) at $\delta = 0$ connecting $\varphi_- = h_r(u_+)$ to $\varphi_+ = h_\ell(u_+)$. This connection exists for some unique value $\theta_F < 0$ of θ. The construction of the fast singular limit again only requires Hypotheses 1 and 2. The slow outer layer is the solution $(u_S(y), h_\ell(u_S(y)), 0, \theta_F)$, where $u_S(y)$ is the solution of

$$u' = -\theta_F g(h_\ell(u),u)$$

satisfying $u_S(0) = u_+$. The singular solution Γ is defined to be the union of all points on the fast layer (for all ξ) with the union of points along the backward slow layer (for $y \leq 0$). Note that θ is included as a dependent variable in the phase space of the wave.

THEOREM 4.3. (a) *Suppose that Hyptotheses* 1–5 *are satisfied. Given* $\eta > 0$, *let* $V_\eta(\Gamma)$ *be an* η*-neighborhood of* Γ. *There exists* $\delta_0 > 0$ *such that for* $\delta < \delta_0$, $S(N \times \{c < 0\}) \subset V_\eta(\Gamma)$, *where* $S(N \times \{c < 0\})$ *denotes the isolated invariant set of* (6) *augmented with the parameter flow* $\dot{\theta} = 0$ *and where* N *is the isolating neighborhood constructed in section* 3.

(b) *Suppose that Hypotheses* 1 *and* 2 *are satisfied. For sufficiently small* η *and* δ, *the perturbed solution in* $V_\eta(\Gamma)$ *is uniquely determined as the transverse intersection of the center-unstable manifold of* P_- *with the center-stable manifold of* P_+ *for the equations in* (6) *augmented with* $\dot{\theta} = 0$. *The connecting solution is a smooth function of* δ *for small* δ, *and in particular, the wave speed* c *has an expansion of the form*

$$c(\delta) = \delta(\theta_F + O(\delta)).$$

Proof. (a) We first obtain another a priori estimate for the scaled wave speed θ. To this end, set $\rho = \dot{\psi}$ and replace ψ with ρ in (6). The $\dot{\psi}$ equation is then replaced by

$$\dot{\rho} = -\theta\rho + F_\varphi(\rho + F)/\theta + F_u(\delta\theta g).$$

If the solution is assumed to lie in $S(N)$, then g and the partials of F are bounded independently of θ and δ. It follows that the ρ equation is of the form

$$\dot{\rho} = (-\theta + O(1/\theta))(\rho + O(\delta)) + O(1/\theta).$$

In order for ρ to remain uniformly bounded, it follows that $|\rho| \leq K(1/|\theta|^2 + \delta)$ for some K independent of both θ and δ. However, the φ equation can be expressed as

$$-\theta\dot{\varphi} = F(\varphi, u) + \rho;$$

for large $|\theta|$ and small δ, it follows that a neighborhood of the right branch $\varphi = h_r(u)$ of $F = 0$ will be positively invariant relative to N. It is therefore impossible to have a connecting orbit from P_- to P_+ for such a θ and δ. There therefore exist $\delta_0 > 0$ and $\theta_1 < 0$ such that when $\delta < \delta_0$, the scaled wave speed θ of any connection in N must lie in the interval $[\theta_1, 0]$. Combining this with the result of Lemma 4.1 shows that $\theta \in [\theta_1, \theta_0]$.

Let $I = [\theta_1, \theta_0]$ and suppose that $\eta > 0$ is given and that

$$(x, \theta) \in (N \times I) \setminus V_\eta(\Gamma).$$

We must show that the solution through this point must leave N in at least one time direction provided that δ is sufficiently small. If this were not the case, the solution through this point would have to be a connecting orbit by Lemma 3.1. We will first show that the wave speed θ of any such connection must approximate the wave speed θ_F of the inner reduced solution. To this end, parametrize the orbit so that $\varphi(0) = \varphi_m$, where φ_m is as in Lemma 3.2. For small δ, the solution is well approximated by setting $\delta = 0$ in (6), in which we set $u \equiv u(0)$. In order for the solution to remain in N, it must approximate the saddle–saddle connection of (6) when $\delta = 0$ on finite ξ intervals. It follows that θ must be near the wave speed $\theta(u(0))$ of the saddle–saddle connection for the $\delta = 0$ system. If $u(0) \leq u_+ - \eta$, then there exists $\xi_0 < 0$ such that the solution is in a small neighborhood of the left slow manifold $\varphi = h_\ell(u(0))$ at ξ_0. It follows from the monotonicity properties of solutions in N that the solution must then remain near the left slow manifold for *all* $\xi \leq \xi_0$. However, since $u(0)$ is strictly less than u_+, it follows that $\dot{u} < 0$ for all $\xi \leq \xi_0$ so that the solution must eventually leave N by having (φ, u) exit M_0 through the bottom of the rectangle. It therefore follows that $u(0)$ must lie within η of u_+ and thus that the wave speed θ must be of order η from θ_F

The proof of (a) is now completed by following the flow through each point in $N \setminus V_\eta(\Gamma)$ and showing that the solution leaves N in at least one time direction. The details of the argument are similar to those used above to show that the wave speed is near θ_F, and we shall only provide a brief outline. Solutions in $S(N)$ for small δ remain outside the middle slow manifold $\varphi = h_*(u)$ since we see from (6) that this is a repelling rest point at $\delta = 0$ so that for small δ, we can construct a negatively invariant region about it. Therefore, a solution through such a point cannot get to P_- in backward time. We now argue as in the previous paragraph that a solution in $S(N)$ must have exactly one transition layer in which it jumps from the right slow manifold to the left slow manifold by approximating the saddle–saddle connection of (6) when $\delta = 0$. Combining these remarks with the estimates for θ and u in the transition layer, it follows that $S(N \times I) \subset V_\eta(\Gamma)$.

(b) By (a), it suffices to describe the wave in the neighborhood $V_\eta(\Gamma)$. To this end, augment (6) with the parameter flow $\dot{\theta} = 0$, let W_δ^{cs} be the center-stable manifold of P_+, and let W_δ^{cu} be the center-unstable manifold of P_- for the augmented flow with $\delta > 0$. Thus W_δ^{cs} is two dimensional and W_δ^{cu} is three dimensional. We note that W_δ^{cs} is foliated by the collection of one-dimensional stable manifolds of P_+ for each θ. Similarly, W_δ^{cu} is foliated by the two-dimensional unstable manifolds of P_- for each θ. These manifolds therefore have limiting configurations W_0^{cs} and W_0^{cu}, respectively, that can be calculated as $\delta \to 0$. The former consists of the collection of one-dimensional stable manifolds $M_s(\theta)$ of P_+ for (6) at $\delta = 0$ for each θ. The latter consists of the two-dimensional center-unstable manifold $M_{cu}(\theta)$ of the continuum of rest points $\varphi = h_r(u)$ for (6) (without the θ equation) for each θ. The tangent space to $M_{cu}(\theta)$ at each one of these rest points consists of a vector tangent to the right slow manifold and a vector in the strongly unstable direction.

The manifolds W_0^{cs} and W_0^{cu} intersect along Γ. We now describe their tangent spaces T^{cs} and T^{cu} at a point Q which lies on the fast layer and close to the "corner point" $P = (u_+, h_r(u_+), 0, \theta_F)$ of the singular limit. The former tangent space at Q consists of a vector $(0, p, q, 0)$, which is close to the strongly unstable direction at the right corner P of Γ, and another vector of the form $(0, r, s, 1)$, where the vectors (p, q) and (r, s) are independent. This is a consequence of the well-known fact that for the scalar bistable travelling wave problem, the saddle–saddle connection occurs as the transverse intersection of center-unstable and center-stable manifolds for the equations augmented with $\dot{\theta} = 0$. The tangent space T^{cu} is spanned by the vectors v_1, v_2, and v_3, where

$$\begin{aligned} v_1 &= (1, h_r'(u_+), 0, 0), \\ v_2 &= (0, p, q, 0), \\ v_3 &= (0, 0, 0, 1). \end{aligned}$$

It immediately follows that these manifolds intersect transversely at Q and therefore everywhere along Γ; thus the perturbed manifolds W_δ^{cs} and W_δ^{cu} also intersect transversely for small δ. The uniqueness of the wave in $V_\eta(\Gamma)$ as well as the differentiable dependence of the wave speed θ on δ immediately follow from this.

The following corollary summarizes the estimates obtained above on the possible wave speeds c for all possible parameter values of α. These estimates are used in [1] to compare the results obtained here from phase-field models with other models of hypercooled solidification.

COROLLARY 4.4. *Under Hypotheses* 1–5*, the wave speed c of any connecting solution of the travelling wave equations in* (1) *which lies interior to the region N constructed in section* 3 *satisfies an estimate of the form*

$$T_0/\alpha < c < T_1/\alpha$$

for some constants $T_0 < T_1 < 0$ depending only on N and for all parameter values $\alpha > 0$.

Proof. The upper bound for c is exactly the result proved in Lemma 4.1 with $T_0 = \theta_0$. A lower bound for c, $\theta_1/\alpha < 0$ also follows from the proof of Lemma 4.1 and Theorem 4.2 for $0 < \alpha < \alpha_0$ for some $\alpha_0 > 0$ since for such an α, the theorem implies that the scaled wave speed $\theta = c\alpha$ must approximate that of the singular limit. The lower bound for c in the parameter range $\alpha > 1/\delta_0$ is proved in a similar manner at the beginning of Theorem 4.3. Finally, in the parameter range $1/\delta_0 < \alpha < \alpha_1$, the estimate follows from the upper and lower bounds for c obtained in Lemma 3.4, which implicitly assumes that α lies in a compact set bounded away from $\alpha = 0$. Combining the estimates above provides the stated result for all $\alpha > 0$.

Remarks. (i) The asymptotics for small α indicate that in this regime, the φ component of the profile should be monotone decreasing in ξ while the u component will either be monotone increasing (in the event that $\{g = 0\}$ is monotone in M_0) or have a unique local minimum at some point. On the other hand, the asymptotics for large α predict that the u-component will be monotone increasing and that the φ-component will have a unique local maximum in this regime. Thus the profile is capable of exhibiting a rich variety of qualitative properties. It is therefore quite plausible that some additional hypotheses are essential in the proof of the global result (Theorem 2.1) linking these two asymptotic regimes.

(ii) Asymptotic expansions to any desired order of accuracy can be obtained in the usual way for both asymptotic regimes. In particular, we derive the first few terms of the expansion for the transition layer in the large α regime. To this end, we expand the solution in powers of δ:

$$\begin{aligned} u &= u_+ + \delta u_1(\xi) + \cdots, \\ \phi &= \varphi_0(\xi) + \delta\varphi_1(\xi) + \cdots, \\ \theta &= \theta_0 + \delta\theta_1 + \cdots, \end{aligned}$$

where $\psi = \varphi'$ has a similar expansion. The zeroth-order terms satisfy the second-order equation

$$\varphi'' + \theta_0\varphi' + F(\varphi_0, u_+) = 0$$

and are determined by taking $\varphi_0(\xi), \theta_0$ to be the solution connecting φ_+ at $+\infty$ to $\varphi_r = h_r(u_+)$ at $-\infty$. In particular, $\theta_0 = \theta_F$. The first-order asymptotics are then determined by the equations

$$\begin{aligned} u_1' &= -\theta_0\gamma(\varphi_+ - \varphi_0(\xi)), \\ \varphi_1'' + \theta_0\varphi_1' + F_\varphi\varphi_1 &= -F_u u_1 - \theta_1\varphi_0', \end{aligned}$$

where the partials of F are evaluated at $(\varphi_0(\xi), u_+)$. The kernel of the adjoint of the linear operator on the left side of the second equation is $\varphi^* = \varphi_0' e^{\theta_0 \xi}$. First, u_1 is determined by integrating the first equation,

$$u_1 = -\theta_0 \gamma \int_\xi^\infty (\varphi_+ - \varphi_0(s)) ds.$$

Since φ_0 decays exponentially to φ_+ (resp. φ_r) at $+\infty$ (resp. $-\infty$), it follows that $u_1(\xi)$ decays exponentially to zero at $+\infty$ and that $u_1(\xi)/\xi$ is asymptotic to $\varphi_+ - \varphi_r$, i.e., u_1 is linear at $-\infty$. It then follows that the integrals in the expression

$$\theta_1 = \frac{-\int_{-\infty}^\infty F_u(\varphi_0(s), u_+) u_1(s) \varphi^*(s)\, ds}{\int_{-\infty}^\infty \varphi^*(s) \varphi_0'(s)\, ds}$$

converge. Hence the first-order asymptotics are completely determined. Clearly, this procedure can now be continued to determine the terms in the asymptotic series to any desired order.

5. Singular limits II: Perturbation by higher-order equations. We finally consider the singular perturbation problem discussed in the introduction,

$$\begin{aligned} \dot u &= -cg(\varphi, u), \\ \dot\varphi_1 &= \varphi_2, \\ \dot\varphi_2 &= \varphi_3, \\ \mu\dot\varphi_3 &= \varphi_4, \\ &\vdots \\ \mu\dot\varphi_{2m-1} &= \varphi_{2m}, \\ \mu\dot\varphi_{2m} &= H(\varphi, u), \end{aligned} \tag{8}$$

where φ now denotes the vector with components φ_i and

$$H(\varphi, u) = -\left(\sum_{n=1}^{m-1} b_n \varphi_{2n+1} + \alpha c \varphi_2 + F(\varphi_1, u)\right) / b_m.$$

We shall fix the parameters in the nonlinearities and α and investigate the behavior of the above system for small μ. Let $\varphi_s = (u, \varphi_1, \varphi_2)$ denote the slow components and $\varphi_f = (\varphi_3, \ldots, \varphi_{2m})$ denote the fast components. We shall locate orbits connecting the rest point $\tilde P_- = (P_-, 0)$ to the rest point $\tilde P_+ = (P_+, 0)$.

The fast–slow structure of the system is more easily seen after the change of independent variables $\xi \to y$, where $y = \mu^{-1}\xi$, and of dependent variables $\varphi_f \to \zeta_f$, where

$$\begin{aligned} \zeta_3 &= \varphi_3 + (c\alpha\varphi_2 + F(\varphi_1, u))/b_2, \\ \zeta_j &= \varphi_j, \quad j = 4, \ldots, 2m. \end{aligned}$$

In the new variables, system (8) can be expressed as

$$\begin{aligned} u' &= -\mu c g(\varphi_1, u), \\ \varphi_1' &= \mu \varphi_2, \end{aligned} \tag{9}$$

$$\begin{aligned}
b_2\varphi_2' &= \mu(-\alpha c\varphi_2 - F(\varphi_1,u) + b_2\zeta_3),\\
\zeta_3' &= \zeta_4 + \mu V(\varphi_s,\zeta_f),\\
\zeta_4' &= \zeta_5,\\
&\vdots\\
\zeta_{2m}' &= -\left(\sum_{n=1}^{m} b_n\zeta_{2n+1}\right)/b_m,
\end{aligned}$$

where V is a smooth function of its arguments, whose precise form is not important here. Formally setting $\mu = 0$, we see that (9) is a singular perturbation of problem (1). This is more easily seen by changing back to the slow (ξ) scaling to obtain

$$\begin{aligned}
\dot{u} &= -cg(\varphi_1,u),\\
\dot{\varphi}_1 &= \varphi_2,\\
b_2\dot{\varphi}_2 &= (-\alpha c\varphi_2 - F(\varphi_1,u) + b_2\zeta_3),\\
\mu\dot{\zeta}_3 &= \zeta_4 + \mu V(\varphi_s,\varphi_f),\\
&\vdots\\
\mu\zeta_{2m} &= -\left(\sum_{n=1}^{m} b_n\zeta_{2n+1}\right)/b_{2m}.
\end{aligned}\tag{10}$$

Note that at $\mu = 0$, the ζ_f equations in (9) form a linear system that is decoupled from φ_s; let J be the coefficient matrix of this $(2m-2)$-dimensional system. It was proved in Lemma 7.1 in [1] that this matrix is always hyperbolic when the coefficients b_n are determined as in [1] by the various moments of the interaction function generating the phase-field model. In particular, J has $(m-1)$ eigenvalues with positive real part and $(m-1)$ eigenvalues with negative real part.

THEOREM 5.1. (a) *Let $V_\eta = \{\zeta_f : |\zeta_f| \le \eta\}$. There exists $\eta > 0$ and $\mu_0 = \mu_0(\eta)$ such that $\tilde{N} = N \times V_\eta$ is an isolating neighborhood for* (9) *for all $\mu \in (0,\mu_0)$ and all $c \in [c_1,c_0]$, where N is the isolating region of Theorem* 2.1 *and c_1 and c_0 are as Lemma* 3.4. *Furthermore, the set of connections from $\tilde{P}_-$ to $\tilde{P}_+$ is empty for $c = c_1, c_0$.*

(b) *The Conley connection index for* (9) *on $\tilde{N} \times [c_1,c_0]$ is the homotopy type of a point so that there exists a connection from $\tilde{P}_-$ to $\tilde{P}_+$ for some $c \in [c_1,c_0]$ and for each $\mu \in (0,\mu_-]$.*

Proof. (a) The boundary of $\tilde{N}$ is $(\partial N \times V_\eta) \cup (N \times \partial V_\eta)$. First, suppose that $\varphi_s \in \partial N$. Since N is a compact isolating region for (1) for each $c \in [c_1,c_0]$, there exists $T > 0$ and $\delta > 0$ such that if N_δ is a δ-neighborhood of N in $\mathbf{R}^3$, then for each $x \in \partial N$, there exists $\xi \in [-T,T]$ such that $x \cdot \xi \in \mathbf{R}^3 \setminus N_\delta$, i.e., the solution through x lies at a distance of at least δ from N at some time ξ that is uniformly bounded from above and below. Now suppose that $(\varphi_s,\zeta_f) \in S(\tilde{N})$ so that $|\zeta_f(\xi)| \le \eta$ for all ξ. It then follows that there exists $K > 0$ such that if $\varphi_s \in \partial N$, then the solution of (10) through (φ_s,ζ_f) satisfies

$$|\varphi_s(\xi) - x(\xi)| \le KT\eta,$$

where $x(\xi)$ is the solution of (1) through φ_s. Now choose η so that $KT\eta \le \delta$. It then follows that such a solution must leave $\tilde{N}$ in time $|\xi| \le T$ so that the solution could not lie in $S(\tilde{N})$.

With η fixed as above, we claim that if $(\varphi_s, \zeta_f) \in N \times \partial V_\eta$, then the solution of (9) through this point exits $\tilde{N}$ in finite time provided that μ is sufficiently small. If this were not the case, the solution would lie in $S(\tilde{N})$ and therefore be uniformly bounded. It follows that the perturbation $\mu V(\varphi_s, \zeta_f)$ in the ζ_3 equation is uniformly of order μ. It therefore follows that if $\bar{\zeta}$ solves $\bar{\zeta}' = J\bar{\zeta}$ with the same initial data as $\zeta_f(y)$, then there exists a constant $K > 0$ depending only on $P > 0$ such that

$$|\bar{\zeta}(y) - \zeta_f(y)| \leq KP\mu$$

for $|y| \leq P$. Since the matrix J is hyperbolic, the neighborhood V_η isolates the origin for the linear flow $\bar{\zeta}' = J\bar{\zeta}$; thus there exists $\varepsilon > 0$ and $P > 0$ such that $\bar{\zeta}(y) \notin V_{\eta+\varepsilon}$ for some $y \in [-P, P]$ whenever $\bar{\zeta}(0) \in \partial V_\eta$. Now choose μ_0 so that $KP\mu_0 < \varepsilon$; it then follows that solutions of (9) which initially lie in $N \times \partial V_\eta$ exit $\tilde{N}$ in time y with $|y| \leq P$. Thus $\tilde{N}$ is isolating for (9) for $\mu \in (0, \mu_0]$.

We next show that the set of connecting orbits from $\tilde{P}_-$ to $\tilde{P}_+$ is empty when $c = c_0, c_1$ for sufficiently small μ. As noted earlier, the matrix J is hyperbolic so that the slow subspace $\{(\varphi_s, 0)\}$ is normally hyperbolic when $\mu = 0$. We may therefore apply Fenichel's theorem [7] to obtain a smooth invariant manifold $\mathcal{M}_\mu = (\varphi_s, \zeta_f(\varphi_s))$, where $\zeta_f(\varphi_s) = O(\mu)$ over compact subsets of the slow subspace. Furthermore, since the rest points $P_\pm$ are hyperbolic for the slow reduced system (1), it follows that the rest points $\tilde{P}_\pm$ are hyperbolic relative to the flow in $\mathcal{M}_\mu$ so that they have stable and unstable manifolds in $\mathcal{M}_\mu$ that closely approximate the stable and unstable manifolds $W_\pm^{u,s}$ of $P_\pm$ of the reduced system (1) for small μ. Furthermore, $\tilde{P}_\pm$ are hyperbolic rest points of (9) so that $\tilde{P}_-$ has an unstable manifold $\tilde{W}_-^u$ and $\tilde{P}_+$ has a stable manifold $\tilde{W}_+^s$ in the full space. By the hyperbolicity of J, these manifolds closely approximate the product manifolds $W_-^u \times U_f$ and $W_+^s \times S_f$, where U_f and S_f are the unstable and stable subspaces of J. By Lemma 3.4, W_-^u and W_+^s do not intersect in N at $c = c_1, c_0$; it follows that $\tilde{W}_-^u$ and $\tilde{W}_+^s$ do not intersect in $\tilde{N}$ for sufficiently small μ for $c = c_1,\ c_0$.

(b) By part (a), the neighborhood $\tilde{N}$ can be used to define a family of connection triples $(\tilde{S}_-, \tilde{S}_+, \tilde{S})$ for system (9) augmented with $c' = 0$ and for all sufficiently small μ. This is achieved by defining $\tilde{S} = S(\tilde{N} \times [c_1, c_0])$ with an analogous definition for $\tilde{S}_\pm$ with N replaced by suitable neighborhoods of the two rest points. The key fact used in part (a) was that the coupling terms $b_2\zeta_3$ and $\mu V(\varphi_s, \zeta_f)$ linking the fast and slow subsystems in (9) are small provided the solution remains in $\tilde{N}$. We now introduce a homotopy of (9) by replacing $b_2\zeta_3$ by $\pi b_2\zeta_3$ and $\mu V(\varphi_s, \zeta_f)$ by $\pi\mu V(\varphi_s, \zeta_f)$, where the homotopy parameter $\pi \in [0, 1]$. Clearly, this only improves the estimates in part (a) so that $\tilde{N}$ generates connection triples $(\tilde{S}_-(\pi), \tilde{S}_+(\pi), \tilde{S}(\pi))$ that are all related by continuation. The connection index $\bar{h}(\pi)$ is therefore independent of π. It immediately follows from the product structure of the system at $\pi = 0$ and the product formula for $\bar{h}$ that $\bar{h}(0)$ is the product of $\bar{h}$ for the three-dimensional slow system (1) with an $(m-1)$-sphere Σ^{m-1}. By the proof of Theorem 2.1, the former index is $[\bar{0}]$, the homotopy type of a point; thus $\bar{h}(1) = \bar{h}(0) = [\bar{0}] \wedge \Sigma^{m-1} = [\bar{0}]$. Since this index is different from the "trivial" index $\Sigma^m \wedge \Sigma^{m+1}$, it follows that $S(\tilde{N})$ contains more than the two rest points $\tilde{P}_\pm$ for some $c \in [c_1, c_0]$. However, any solution which remains in $\tilde{N}$ for all ξ must lie in the invariant manifold $\mathcal{M}_\mu$. As in Lemma 3.1, it follows that any such solution must be a connecting orbit from $\tilde{P}_-$ to $\tilde{P}_+$.

As before, finer information about the wave can be obtained when α is either small or large. The following theorem is the natural generalization of the invariant-manifolds approach used by Gardner and Jones in [11] for the (scalar) higher-order phase-field equation in which the temperature is constant.

THEOREM 5.2. *Let the heteroclinic solution of* (1) *obtained when α is either small or large be denoted by*

$$(u_r(\xi,\alpha),\varphi_r(\xi,\alpha),\psi_r(\xi,\alpha),c_r(\alpha)).$$

For fixed α, there exists μ_0 such that for $\mu \in (0,\mu_0]$, the solution of (10) (*augmented with* $c'=0$) *is unique and can be expanded in a series of the form*

$$\begin{aligned}\varphi_s(\xi,\mu) &= (u_r(\xi,\alpha),\varphi_r(\xi,\alpha),\psi_r(\xi,\alpha)) + O(\mu),\\ \zeta_f(\xi,\mu) &= O(\mu),\\ c(\mu) &= c_r(\alpha) + O(\mu).\end{aligned}$$

Proof. The proof in either case is similar to those using Fenichel's theorem in the previous section. We shall therefore only give a brief outline in the case of small $\delta = 1/\alpha$. Let $c = \delta\theta$ and $\alpha c = \theta$ in (9) and augment this system with the parameter flow $\theta' = 0$. The augmented system has a four-dimensional manifold of rest points $\mathcal{M}_0$ at $\mu = 0$, namely, $\zeta_f = 0$, and by the hyperbolicity of the matrix J proved in Lemma 7.1 of [1], this manifold is normally hyperbolic. By [7], there exists a smooth invariant manifold $\mathcal{M}_\mu$ of the augmented system for small μ, and the flow on $\mathcal{M}_\mu$ is of the form

$$\begin{aligned}\theta' &= 0,\\ u' &= -\delta\theta g(\varphi_1,u) + O(\mu),\\ \varphi_1' &= \varphi_2 + O(\mu),\\ \varphi_2' &= -\theta\varphi_2 - F(\varphi_1,u) + O(\mu).\end{aligned}\tag{11}$$

In Theorem 4.3, it was proved that at $\mu = 0$, the above system has a connecting solution from P_- to P_+ for some θ and for small δ which lies in the transverse intersection of the center-unstable manifold of the former rest point and the center-stable manifold of the latter. It is easily seen from (8) that the perturbed system has the same rest points (with $\zeta_f = 0$) so that (11) has rest points P_- and P_+ with invariant manifolds W^{cu} and W^{cs} which smoothly approximate those of the reduced equations with $\mu = 0$. Since the reduced manifolds intersect transversely, this intersection must persist for small μ.

We note that all that was used in the above argument was the normal hyperbolicity of $\mathcal{M}_0$ and the transverse intersection of the two invariant manifolds of the reduced flow. The same argument therefore applies to the small α regime as well.

6. Numerical results.

We conclude with the results of some numerical experiments involving the second-order phase-field system

$$\begin{aligned}(u + \lambda\varphi - a\varphi^2)_t &= u_{xx},\\ \alpha\varepsilon^2\varphi_t &= \varepsilon^2\varphi_{xx} + b(\varphi - \varphi^3) + u\end{aligned}$$

in one space-variable x on a large finite interval $[-50, 50]$ with homogeneous Neumann boundary conditions and step-function initial data

$$(\varphi(x,0),u(x,0)) = \begin{cases} \varphi_-, & x < 20,\\ \varphi_+(\lambda), & x > 20.\end{cases}$$

Note that the equations are parameterized in a slightly different manner than the system described in the introduction. In particular, we take $W(\varphi) = \varphi - B\varphi^2$ in the equations described in the introduction; then those equations are equivalent to the above with $a = \lambda B$. In most situations, the interval above was large enough that the boundary did not appear to affect the transient behavior of the solution on time intervals large enough for travelling waves to be observed. The equations were integrated numerically with the MOL1D package which is based on the method of lines and a GEAR program to solve the associated system of ODEs. Solutions were calculated on several different meshes to check for accuracy. In all but one case (Figure 9 below), a mesh of 300 points appeared to be fine enough to give reliable results in that the wave speeds and profiles calculated on different grids matched to a high degree of accuracy.

Experiments were performed in broad ranges of the various parameters. In the following, the null sets of the nonlinearities, $F = 0$ and $g = 0$, the u-component, and the φ-component of the solution are depicted in the first, second, and third figures of each sequence. To begin with, parameters were set so as to be consistent with the hypotheses of Theorem 2.1 to test both the theorem and the numerical code. In Figure 5, $\lambda = .06$ and $a = 0$ so that $g = 0$ is close to a horizontal line and the wave is close to the scalar wave with constant u. Note that both components of the solution are monotone. In Figure 6, λ was increased so that the hypotheses of Theorem 2.1 fail to be satisfied. For example, it can be seen from Figure 6a that it is impossible to construct a rectangle M_0 which intersects $F = 0$ and $g = 0$ in the manner required in Hypothesis 3; it is also impossible to construct the rectangle M_- about P_-. However, travelling waves were still observed in the solution, indicating that the hypotheses of Theorem 2.1, while sufficient, are far from necessary conditions for the existence of the profile. The nonlinearities in Figure 7 are still consistent with the hypotheses of Theorem 2.1; however, $W(\varphi)$ is quadratic. Note that the u-component now has a local minimum while φ is still monotone decreasing. This type of profile was predicted by the asymptotic results of Theorem 4.2 with small α, although α in this particular experiment was $O(1)$. In Figure 8, the coefficient of the quadratic term in W was sufficiently increased so that it is again impossible to construct the rectangles M_0 and $M_\pm$ as required in Hypotheses 3 and 4 of Theorem 2.1. Nevertheless, the system still supports travelling waves.

Finally, in Figure 9, we illustrate another interesting limiting regime in which u_+ is near zero. In this case, the solution has a structure similar to that of the solutions constructed in Theorem 4.3, in which φ has a transition layer and the temperature u is monotone and has a “corner” near the point where the phase variable has a transition layer. This is a different singular limit problem than that of Theorem 4.3, wherein $\delta = 1/\alpha$ was assumed to be small. Here the small parameter is the scaled wave speed θ in (6), which is determined by the heteroclinic solution of the last two equations in (6) when the temperature is the constant $u = u_+$. This singular limit problem does not fit immediately into the framework of Theorem 4.3, and it requires a separate analysis. Nevertheless, it seems likely that an existence theorem along the lines of Theorem 4.3 could be obtained for small $u_+ < 0$.

It is interesting to note that we were forced to consider a much finer spatial mesh in order to observe the waves in Figure 9. Coarser meshes seemed to give rise to Hopf bifurcations in the form of periodic hot spots appearing in the temperature profile. This must be regarded as a numerical artifact since the phenomenon disappeared with mesh refinement.

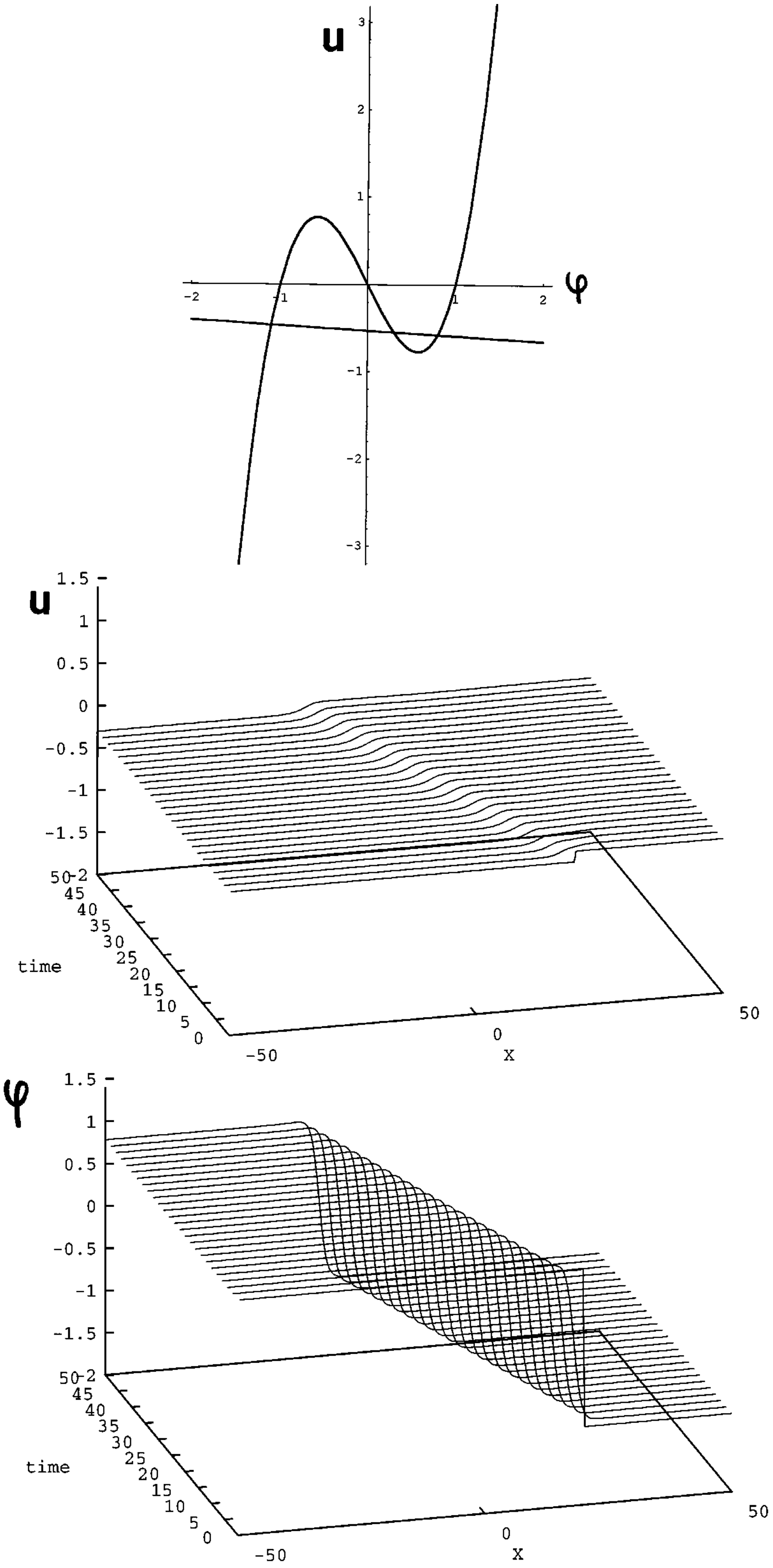


FIG. 5. $(\varphi, u)_- = (.8, -.29)$, $(\varphi, u)_+ = (-1.1, -.23)$, $\lambda = .06$, $a = 0$, $b = 2$, $\varepsilon = 1$, $\alpha = 1$.

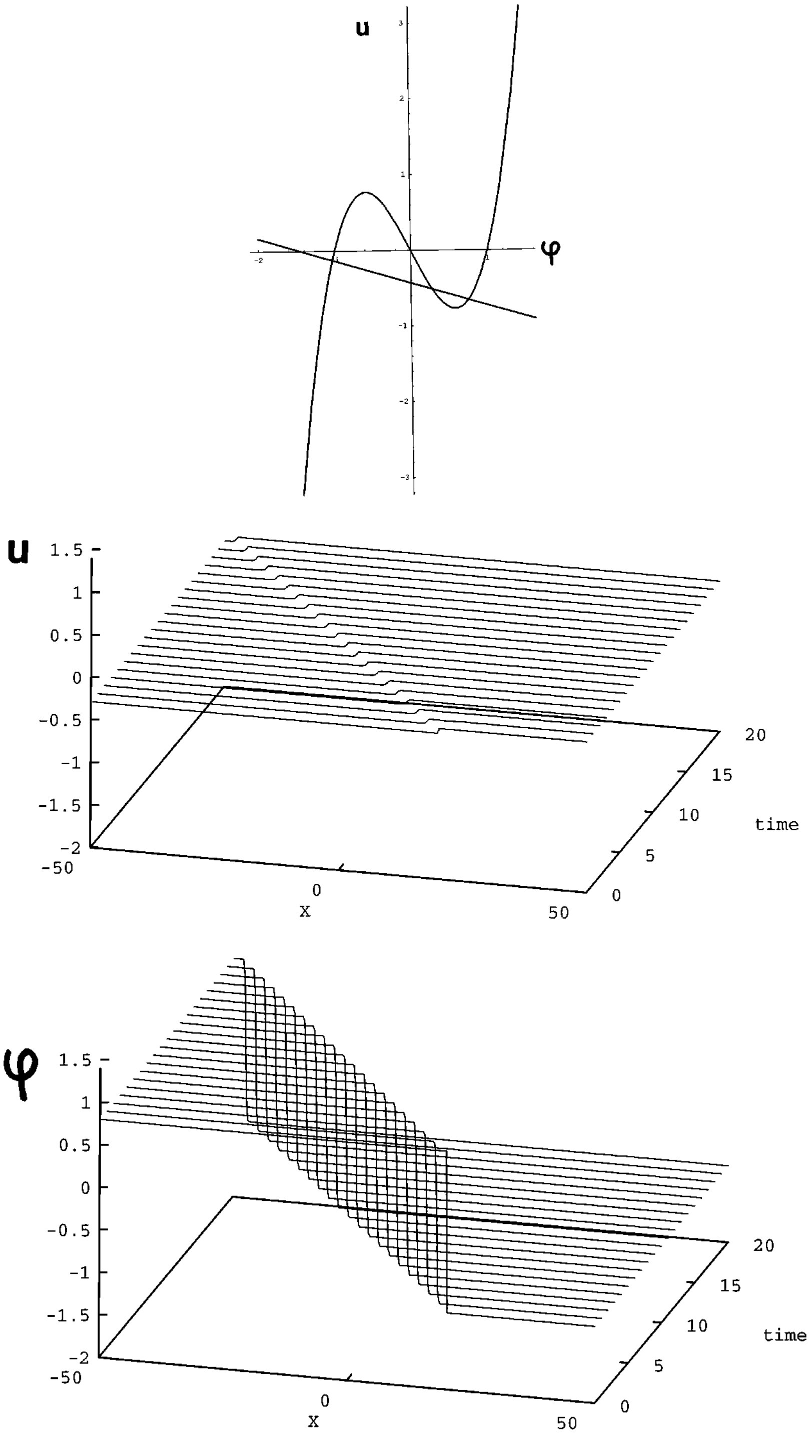


FIG. 6. $(\varphi, u)_- = (.75, -.66)$, $(\varphi, u)_+ = (-1.03, -.06)$, $\lambda = .3$, $a = 0$, $b = 2$, $\varepsilon = 1$, $\alpha = 1$.

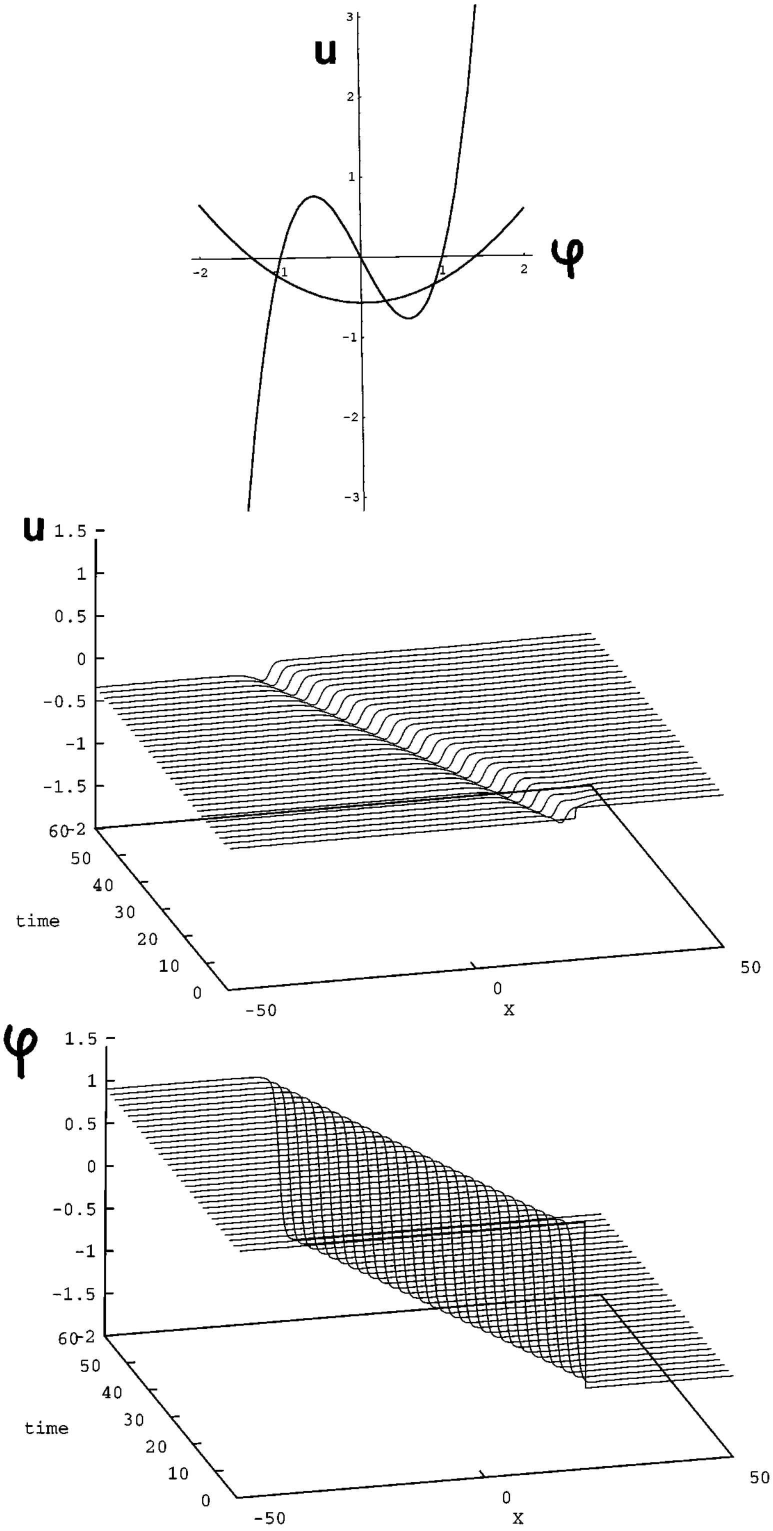


FIG. 7. $(\varphi, u)_- = (.9, -.34)$, $(\varphi, u)_+ = (-1.05, -.21)$, $\lambda = .02$, $a = .3$, $b = 2$, $\varepsilon = 1$, $\alpha = 1$.

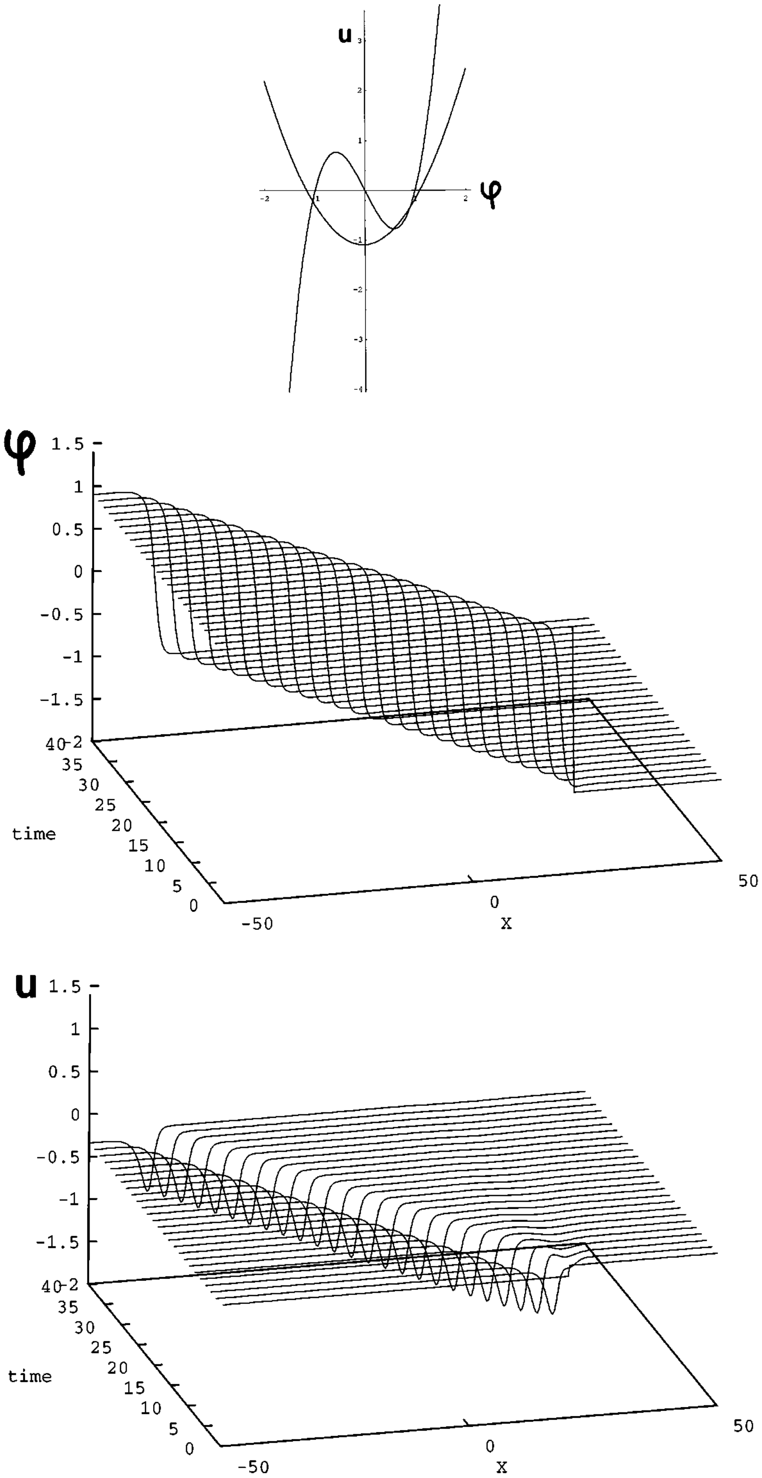


FIG. 8. $(\varphi, u)_- = (.9, -.34)$, $(\varphi, u)_+ = (-1.05, -.21)$, $\lambda = -.06$, $a = .85$, $b = 2$, $\varepsilon = 1$, $\alpha = 1$.

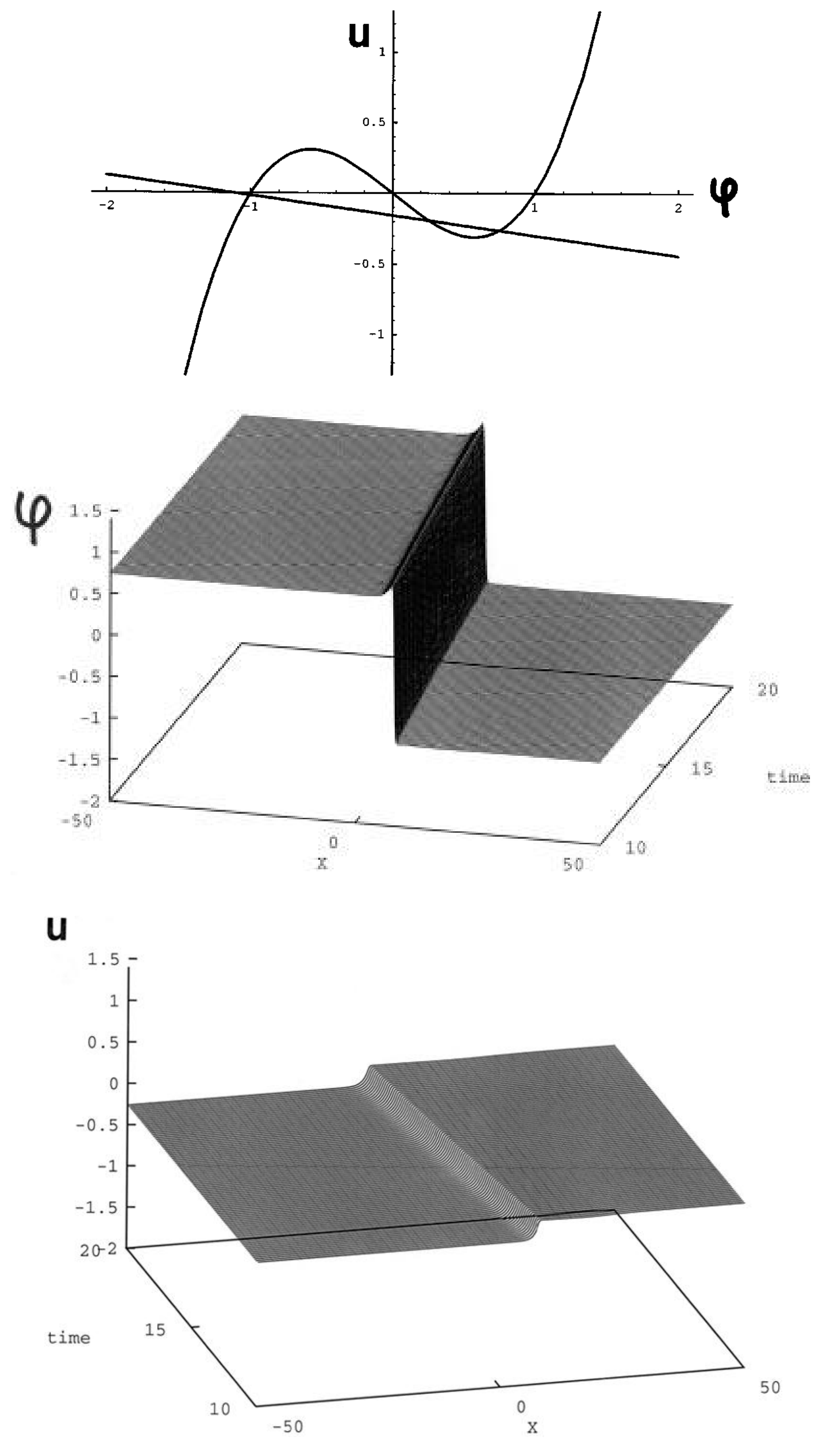


FIG. 9. $(\varphi, u)- = (.75 - .66)$, $(\varphi, u)_+ = (-1.01, -.008)$, $\lambda = .14$, $b = .8$, $a = 0$, $b = .8$, $1/\varepsilon^2 = 55$, $\alpha = 1$.

We also attempted to calculate the singular solutions of Theorem 4.3 with α large and u_+ negative and bounded away from zero. The numerical solutions did not produce travelling wave profiles consistent with the singular limit structure of Theorem 4.3. The temperature profile appeared more like the solution of a linear diffusion equation with step-function initial data, while the phase variable appeared to have a nearly stationary transition layer. In particular, the numerical solutions did not appear to converge to travelling waves. Whether the source of this is numerical or whether it points to some underlying instability of the travelling wave profiles of Theorem 4.3 is at this point unclear.

REFERENCES

[1] P. Bates, P. Fife, R. Gardner, and C. Jones, *Phase field models for hypercooled solidification*, Phys. D., to appear.

[2] G. Caginalp, *The role of microscopic anisotropy in the macroscopic behavior of a phase boundary*, Ann. Phys., 172 (1986), pp. 136–155.

[3] G. Caginalp and P. Fife, *Higher order phase field models and detailed anisotropy*, Phys. Rev. B, 34 (1986), pp. 4940–4943.

[4] G. Caginalp and P. Fife, *Dynamics of layered interfaces arising form phase boundaries*, SIAM J. Appl. Math., 48 (1988), pp. 506–518.

[5] G. Caginalp and Y. Nishiura, *The existence of travelling waves for phase field equations and convergence to sharp interface models in the singular limit*, Quart. Appl. Math., 49 (1991), pp. 147–162.

[6] C. Conley and R. A. Gardner, *An application of the generalized Morse index to a competitive diffusion-reaction model*, Indiana Univ. Math. J., 33 (1984), pp. 319–343.

[7] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[8] P. Fife and O. Penrose, *Interfacial dynamics for theromodynamically consistent phase-field models with nonconserved order parameter*, Electronic J. Differential Equations, 1995, pp. 1–49.

[9] P. C. Fife, *Dynamics of internal layers and diffusive interfaces*, CBMS–NSF Regional Conference Series in Applied Mathematics 53, SIAM, Philadelphia, 1988.

[10] R. A. Gardner, *Existence of travelling wave solutions of predator-prey systems via the connection index*, SIAM J. Appl. Math., 44 (1984), pp. 56–79.

[11] R. A. Gardner and C. Jones, *Stability of travelling waves of diffusive predator prey systems*, Trans. Amer. Math. Soc., 327 (1991), pp. 465–524.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 94–108, January 1997

COMPRESSIBLE NAVIER–STOKES EQUATIONS IN A BOUNDED DOMAIN WITH INFLOW BOUNDARY CONDITION[*]

JAE RYONG KWEON[†] AND R. BRUCE KELLOGG[‡]

Abstract. In this paper, we study the barotropic compressible Navier–Stokes equations in a bounded plane domain Ω. Nonzero velocities are prescribed on the boundary of Ω, and the density is prescribed on that part of the boundary corresponding to entering velocity. This causes a weak singularity in the solution at the junction of incoming and outgoing flows. We prove the existence of the solution $(\mathbf{u}, p)$ of the system

$$\begin{cases} -\mu\Delta\mathbf{u} - \nu\nabla\mathrm{div}\,\mathbf{u} + \rho(p)(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = 0 \quad \text{in } \Omega, \\ \mathrm{div}(\rho\mathbf{u}) = 0 \quad \text{in } \Omega, \\ \mathbf{u} = \mathbf{u}_0(x,y) \quad \text{on } \Gamma, \\ p = p_0(x,y) \quad \text{on } \Gamma_{\mathrm{in}} \end{cases}$$

in the Sobolev space $H^{2,q} \times H^{1,q}(2 < q < 3)$. The proof follows from an analysis of the linearized problem and a fixed-point argument.

Key words. Navier–Stokes equations, singularities, compressible viscous flows, fixed-point arguments

AMS subject classifications. 35Q30, 76N10

PII. S0036141095284254

1. Introduction and main results.

The steady-state barotropic compressible Navier–Stokes equations are a system of PDEs of mixed type; the momentum equations form an elliptic subsystem in the velocity components, and the continuity equation is a hyperbolic equation in the density. Because, as a rule, the flow is zero on the boundary of a region, boundary value problems for the Navier–Stokes system are generally considered with the condition that the velocity components vanish on the boundary of the region. It is at least of mathematical interest and possibly of physical interest to consider boundary value problems in which the velocity components assume specified nonzero values on the boundary. Since the velocity field gives the characteristic directions for the continuity equation, values of the density must be specified on those portions of the boundary where the specified velocity vector points into the region. In this paper, we shall discuss the resulting boundary value problem in a special case. Particular attention will be paid to the boundary points where the velocity vector is tangent to the boundary since these give rise to singularities in the solution. We study the system

$$\begin{cases} -\mu\Delta\mathbf{u} - \nu\nabla\mathrm{div}\,\mathbf{u} + \rho(p)(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = 0 \quad \text{in } \Omega, \\ \mathrm{div}(\rho\mathbf{u}) = 0 \quad \text{in } \Omega, \\ \mathbf{u} = \mathbf{u}_0(x,y) \quad \text{on } \Gamma, \\ p = p_0(x,y) \quad \text{on } \Gamma_{\mathrm{in}}. \end{cases} \tag{1.1}$$

Here Ω is a bounded open set in R^2 with smooth boundary Γ, $\mathbf{u} = [u, v]$ is the velocity vector, p is the pressure, $\rho = \rho(p)$ is the density, μ is the viscosity constant, and ν

[*] Received by the editors April 5, 1995; accepted for publication September 13, 1995.
http://www.siam.org/journals/sima/28-1/28425.html

[†] Department of Mathematics, University of Maryland, College Park, MD 20742 (kweon@math.snu.ac.kr).

[‡] Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (kellogg@olgao.umd.edu).

is the bulk viscosity constant. We assume that $\mu > 0$ and $\nu > -\mu$. The functions $\mathbf{u}_0(x, y)$ and $p_0(x, y)$ that give the boundary data are assumed to be smooth functions on the closure of Ω. The incoming and outgoing portions of the boundary, Γ_{in} and Γ_{out}, are defined by

$$\begin{aligned}\Gamma_{\text{in}} &= \{(x, y) \in \Gamma : \mathbf{u}_0 \cdot \mathbf{n} < 0\},\\ \Gamma_{\text{out}} &= \{(x, y) \in \Gamma : \mathbf{u}_0 \cdot \mathbf{n} \geq 0\},\end{aligned}$$

where $\mathbf{n} = [n_1, n_2]$ denotes the unit outward-pointing normal to Γ. In fact, to make the construction simpler, we assume throughout the paper that $v_0(x, y) \equiv 0$, $u_0(x, y) \geq C_0 > 0$.

Let $\Gamma_* \subset \Gamma$ be the set of the points at which $\mathbf{u}_0 \cdot \mathbf{n}$ is zero. We shall further simplify the situation by imposing the following condition.

Condition A. Γ_* consists of two points (x_*, y_*) and (x^*, y^*).

Again, Condition A is not essential for our results but serves to remove unessential details. With Condition A, the boundary Γ is divided into two connected arcs, the arc Γ_{in} for which $\mathbf{u}_0 \cdot \mathbf{n} < 0$ and the arc Γ_{out} for which $\mathbf{u}_0 \cdot \mathbf{n} \geq 0$. There are two increasing functions $\delta_{\pm}(y)$, defined for $y_* \leq y \leq y^*$, such that

$$\begin{aligned}\Gamma_{\text{in}} &= \{(\delta_-(y), y) \in \Gamma : y_* \leq y \leq y^*\},\\ \Gamma_{\text{out}} &= \{(\delta_+(y), y) \in \Gamma : y_* \leq y \leq y^*\},\\ \Omega &= \{(x, y) : \delta_-(y) < x < \delta_+(y), y_* \leq y \leq y^*\}.\end{aligned}$$

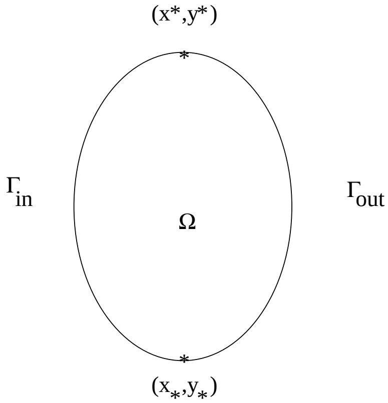


FIG. 1

(See Figure 1.) On Γ_{in}, we have

$$\mathbf{n}(\delta_-(y), y) = \frac{1}{\sqrt{1 + \delta_-'(y)^2}}[-1, \delta_-'(y)]. \tag{1.2}$$

Also, in a neighborhood of (x^*, y^*), Γ is the graph of a function $y = \epsilon_+(x)$, and in a neighborhood of (x_*, y_*), Γ is the graph of a function $y = \epsilon_-(x)$. Near (x^*, y^*) or (x_*, y_*), we have

$$\mathbf{n}(x, \epsilon_{\pm}(x)) = \frac{1}{\sqrt{1 + \epsilon_{\pm}'(x)^2}}[-\epsilon_{\pm}'(x), \pm 1].$$

Near these two points, we define $\mu(x) = (\mathbf{u}_0 \cdot \mathbf{n})(x, \epsilon(x))$, where ϵ denotes ϵ_- or ϵ_+. For simplicity, we will use ϵ to stand for ϵ_- or ϵ_+, respectively.

We will also use the following nondegeneracy assumption.

Condition B. $\mu'(x_*) \neq 0$ and $\mu'(x^*) \neq 0$.

With our assumptions, Condition B is equivalent to the condition that Γ has nonzero curvature at (x_*, y_*) and (x^*, y^*).

We define the distance from Γ_* to any point on the boundary Γ_{in}:

$$d(y) \equiv \min\left\{\sqrt{(y-y^*)^2 + (\delta_-(y) - x^*)^2}, \sqrt{(y-y_*)^2 + (\delta_-(y) - x_*)^2}\right\}. \tag{1.3}$$

Remark. Assume that Condition B holds. Then there are positive constants C_1 and C_2 such that

$$C_1 d(y) \leq \min\{|y-y^*|^{\frac{1}{2}}, |y-y_*|^{\frac{1}{2}}\} \leq C_2 d(y). \tag{1.4}$$

Proof. If $y = \epsilon(x)$, then $x = \delta_-(y)$ by the description of the boundary Γ_{in}. Thus $y - y^* = \epsilon'(\xi)(x - x^*)$ for some $\xi \in (x, x^*)$. Since $\mu(x^*) = 0$ and $\mu'(x^*) \neq 0$, we get

$$\epsilon'(\xi) = \int_{x^*}^{\xi} \epsilon''(s)ds \sim C(\xi - x^*) \quad \text{near } x^*.$$

Hence $y - y^* \sim C(x - x^*)^2 = C(\delta_-(y) - x^*)^2$ near x^* and for some C. However, $d(y) = |\delta_-(y) - x^*|\sqrt{1 + \epsilon'(\xi)^2}$ for some ξ. Thus we easily get $C_1 d(y) \leq \sqrt{y^* - y} \leq C_2 d(y)$ for some positive constants C_1 and C_2, and similarly for $\mu'(x_*) \neq 0$. □

The following lemma describes the behavior of the singularity near Γ_*.

LEMMA 1.1. *Let $\mathbf{u}_0 = [u_0, 0]$ be a given smooth vector field with $u_0 \geq C_0 > 0$. Assume that Condition* B *holds. Then there exist positive numbers m and M such that $md(y) \leq |\mu(\delta_-(y))| \leq Md(y)$ near Γ_*.*

Proof. Let y be near y^*, and let $x = \delta_-(y)$, so $y = \epsilon(x)$. Since $\mu(x^*) = 0$,

$$\begin{aligned}\mu(\delta_-(y)) &= (\mathbf{u}_0 \cdot \mathbf{n})(\delta_-(y), \epsilon(\delta_-(y))) \\ &= (\mu \circ \delta_-)'(\zeta)(\delta_-(y) - x^*)\end{aligned}$$

for some ζ with $y \leq \zeta \leq y^*$. Since $(\mu \circ \delta_-)'(y)$ is continuous in y and $\mu'(\delta_-(y^*)) > 0$, there are positive numbers m_1 and M_1 such that $m_1 \leq \mu'(\delta_-(y)) \leq M_1$ near y^*. However, by (1.3), $d(y) = |\delta_-(y) - x^*|\sqrt{\epsilon'(\delta_-(\zeta))^2 + 1}$ for some $\zeta \in [\bar{y}, y^*]$. Hence

$$\mu(\delta_-(y)) = \frac{(\mu \circ \delta_-)'(\zeta)}{\sqrt{1 + \epsilon'(\delta_-(\zeta))^2}} d(y)$$

for some $\zeta \in [\bar{y}, y^*]$. Now if $\epsilon'(x)$ is bounded near $x = x^*$, then

$$m \leq \frac{(\mu \circ \delta_-)'(\zeta)}{\sqrt{1 + \epsilon'(\delta_-(\zeta))^2}} \leq M \quad \text{near } y^*$$

for some positive constants m and M. A similar argument holds in a neighborhood of (x_*, y_*). □

As an example, let $\mathbf{u}_0$=(1, 0) and consider the domain $\Omega = \{(x,y) \in R^2 : |y| < 1 - x^2, \ -1 < x < 1\}$. Then if $y \geq 0$, $\delta_-(y) = -\sqrt{1-y}$, $\epsilon_+(x) = 1 - x^2$, and $y = \epsilon_+(\delta_-(y))$. Now (0, 1) is the singularity point. Hence $d(y) = \sqrt{(1-y)^2 + \delta_-(y)^2}$ and $|\mathbf{u}_0 \cdot \mathbf{n}| = 2\sqrt{1-y} = 2d(y)/\sqrt{2-y}$, where $\mathbf{n} = (2x, 1)$. Hence $d(y)/3 \leq |(\mu \circ \delta_-)(y)| \leq d(y)$, and similarly for $y < 0$.

We now explain our reason for taking $\mathbf{u}_0$ in the form $\mathbf{u}_0 = [u_0, 0]$. Let $\mathbf{u}_0 = [u_0, v_0]$ with v_0 not necessarily 0 be a given smooth vector field on the closure of Ω with $u_0 \geq C_0 > 0$—say $\mathbf{u}_0 \in C^\infty(\bar{\Omega})$. Consider the function $k(x, \bar{y})$ generated by the vector $\mathbf{u}_0$ and defined by the following first-order ODE: for each fixed $\bar{y}$,

$$\begin{aligned} k_x(x,\bar{y}) &= u_0^{-1} v_0(x, k(x,\bar{y})), \\ k(\delta_-(\bar{y}), \bar{y}) &= \bar{y}. \end{aligned} \tag{1.5}$$

We also consider the streamlines generated by this function, that is, the points $(x, k(x, \bar{y}))$. The solution k of (1.5) is given as follows: setting $U_0 = u_0^{-1} v_0$,

$$k(x,\bar{y}) = \int_{\delta_-(\bar{y})}^{x} U_0(s, k(s,\bar{y}))ds + \bar{y} \quad \text{for all x.} \tag{1.6}$$

Here we can observe that the function $k(x, \bar{y})$ is strictly increasing in $\bar{y}$. Now we set $y = k(x, \bar{y})$. Thus $\bar{y} = \psi(x, y)$ for some function ψ. Hence equation (1.6) can be written as

$$y = \int_{\delta_-(\psi(x,y))}^{x} U_0(s, k(s, \psi(x,y)))ds + \psi(x,y). \tag{1.7}$$

Now we want to compute the first and second derivatives of $\delta_-(\psi(x, y))$. First, we have

$$\nabla \delta_-(\psi(x,y)) = \delta'_-(\psi(x,y)) \nabla \psi(x,y), \tag{1.8}$$

$$\nabla^2 \delta_-(\psi(x,y)) = \delta''_-(\psi(x,y))(\nabla\psi)^2 + \delta'_-(\psi(x,y))\nabla^2\psi. \tag{1.9}$$

We compute $\nabla\psi(x, y)$. Differentiating both sides of (1.7) with respect to x and y, respectively, we get

$$\psi_x(x,y) = (\mathbf{u}_0 \cdot \mathbf{n})^{-1} \left\{ -v_0 - u_0 \int_{\delta_-(\psi(x,y)}^{x} \frac{\partial}{\partial x}[U_0(s, k(s, \psi(x,y)))]\mathrm{ds} \right\}, \tag{1.10}$$

$$\psi_y(x,y) = (\mathbf{u}_0 \cdot \mathbf{n})^{-1} \left\{ v_0 - u_0 \int_{\delta_-(\psi(x,y)}^{x} \frac{\partial}{\partial y}[U_0(s, k(s, \psi(x,y)))]\mathrm{ds} \right\}, \tag{1.11}$$

where $\mathbf{n}$ is the unit normal vector defined in (1.2). Since $\mathbf{u}_0 \cdot \mathbf{n}$ is zero at the points of Γ_*, ψ_x and ψ_y have singularities there. If $v_0 = 0$, $\delta'_-(y)$ is ∞ on Γ_* . In either case, the behavior of the singularities will be the same, and there is no mathematical difference between the two cases. According to this observation, throughout this paper, we choose $\mathbf{u}_0 = [u_0, 0]$ with $u_0 > 0$. With this condition, the flow enters the region from the left. Around Γ_*, the streamlines in Ω are almost parallel to the vector $\mathbf{u}_0$ because $\mathbf{u}|_\Gamma = \mathbf{u}_0$.

In this paper, we obtain the following main results. The proofs are given in section 2.

THEOREM 1.1. *Suppose that* $\mathbf{u}_0 \in H^{2,q}(\Omega)$, $p_0 \in H^{1,q}(\Omega)$, *and* $2 < q < 3$. *Then there is a constant* μ^* *depending on* $\|\mathbf{u}_0\|_{2,q} + \|p_0\|_{1,q}$ *such that if* $\mu \geq \mu^*$, *then problem* (1.1) *has a unique solution* $(u, v, p) \in H^{2,q} \times H^{2,q} \times H^{1,q}$ *with the following estimate:*

$$\|u - u_0\|_{2,q} + \|v\|_{2,q} + \|p - p_0\|_{1,q} \leq C_1,$$

where $C_1 = C(\Omega, \mu, \nu, C_0, \mu^*, |\kappa_0|_{1,\infty}, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$.

The condition that μ is large means that the Reynolds number of the flow is small, so the flow in this domain is a laminar flow.

The next theorem does not have a restriction on μ. Instead, it is required that the functions $\mathbf{u}_0$ and p_0 are almost constant. The proof of Theorem 1.2 is also given in section 2.

THEOREM 1.2. *Let $\mathbf{u}_0 \in H^{2,q}$, $p_0 \in H^{1,q}$, and $2 < q < 3$. For any constant C_1, there is a constant C_2 such that if $\|\mathbf{u}_0\|_{2,q} + \|p_0\|_{1,q} \le C_1$ and $\|\nabla \mathbf{u}_0\|_{1,q} + \|\nabla p_0\|_{1,q} \le C_2$, then there is a unique solution $(u, v, p) \in H^{2,q} \times H^{2,q} \times H^{1,q}$ of system* (1.1) *with the inequality*

$$\|u - u_0\|_{2,q} + \|v\|_{2,q} + \|p - p_0\|_{1,q} \le C_3,$$

where $C_3 = C(\Omega, \mu, \nu, C_0, C_1, C_2, \|\kappa_0\|_{1,q}, |\rho|_\infty, |\tau|_{1,\infty}, |\bar{\rho}|_\infty)$.

The proof of the theorems above consist of formulating problem (1.1) as a mapping on a certain Banach space in such a way that the solution to (1.1) is a fixed point of the mapping. The Schauder fixed-point theory is then used to establish the existence of a fixed point. We use the Banach space of pairs $(\mathbf{u}, p) \in (H^{2,q}(\Omega))^2 \times H^{1,q}$. The index q is chosen in the open interval $(2, 3)$. We choose $q > 2$ so that the Sobolev imbedding theorem can be applied to guarantee that the density $\rho(p)$ is well defined. We require that $q < 3$ to handle the singularities in the solution that occur on Γ_*.

In our analysis, it is convenient to use the deviation from the boundary values as dependent variables. We therefore define new dependent variables $\bar{\mathbf{u}} = \mathbf{u} - \mathbf{u}_0$ and $\bar{p} = p - p_0$. We then obtain for the first equation in (1.1) that

$$-\mu\Delta(\bar{\mathbf{u}} + \mathbf{u}_0) - \nu\nabla \mathrm{div}(\bar{\mathbf{u}} + \mathbf{u}_0) + \rho[(\bar{\mathbf{u}} + \mathbf{u}_0) \cdot \nabla](\bar{\mathbf{u}} + \mathbf{u}_0) + \nabla(\bar{p} + p_0) = 0,$$

and rearranging this equation, we get

$$-\mu\Delta\bar{\mathbf{u}} - \nu\nabla \mathrm{div}\, \bar{\mathbf{u}} + \nabla\bar{p} + \rho\{[(\bar{\mathbf{u}} + \mathbf{u}_0) \cdot \nabla]\bar{\mathbf{u}} + (\bar{\mathbf{u}} \cdot \nabla)\mathbf{u}_0\} = \mathbf{f}, \tag{1.12}$$

where $\mathbf{f} = \mu\Delta\mathbf{u}_0 + \nu\nabla \mathrm{div}\, \mathbf{u}_0 - \nabla p_0 - \rho(\mathbf{u}_0 \cdot \nabla)\mathbf{u}_0$. By the relation $\rho = \rho(p)$, $\mathrm{div}(\rho\mathbf{u}) = \rho(p)\mathrm{div}\mathbf{u} + \rho^{'}(p)\mathbf{u} \cdot \nabla p = 0$, and $\kappa(p)\mathrm{div}\, \mathbf{u} + \mathbf{u} \cdot \nabla p = 0$, where $\kappa(p) = \rho(p)\rho^{'}(p)^{-1}$. Hence the second equation in (1.1) becomes

$$\kappa(\bar{p} + p_0)\mathrm{div}\, \bar{\mathbf{u}} + (\bar{\mathbf{u}} + \mathbf{u}_0) \cdot \nabla p + \bar{\mathbf{u}} \cdot \nabla p_0 = g, \tag{1.13}$$

where $g = -\kappa(\bar{p} + p_0)\mathrm{div}\, \mathbf{u}_0 - \mathbf{u}_0 \cdot \nabla p_0$. Here we define $\tau(p)$ and $\bar{\rho}(p)$ as follows:

$$\tau(p) \equiv \frac{\kappa(p) - \kappa(p_0)}{p - p_0} \quad \text{for } p \ne p_0, \quad \tau(p_0) = \kappa'(p_0), \tag{1.14}$$

$$\bar{\rho}(p) \equiv \frac{\rho(p) - \rho(p_0)}{p - p_0} \quad \text{for } p \ne p_0, \quad \bar{\rho}(p_0) = \rho'(p_0). \tag{1.15}$$

Thus $\kappa(p) = \kappa(p_0) + \tau(p)(p - p_0)$ and $\rho(p) = \rho(p_0) + \bar{\rho}(p)(p - p_0)$. Now we join (1.12) and (1.13) and replace $\bar{\mathbf{u}}$ and $\bar{p}$ by $\mathbf{u}$ and p for convenience. Then we get

$$\left\{\begin{array}{l} -\mu\Delta\mathbf{u} - \nu\nabla \mathrm{div}\, \mathbf{u} + \nabla p + \rho\{[(\mathbf{u} + \mathbf{u}_0) \cdot \nabla]\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u}_0\} = \mathbf{f} \quad \text{in } \Omega, \\ \kappa(p + p_0)\mathrm{div}\, \mathbf{u} + (\mathbf{u} + \mathbf{u}_0) \cdot \nabla p + \mathbf{u} \cdot \nabla p_0 = g \quad \text{in } \Omega, \\ \mathbf{u} = 0 \quad \text{on } \partial\Omega, \\ p = 0 \quad \text{on } \Gamma_{\text{in}}, \end{array}\right. \tag{1.16}$$

where $\mathbf{f}$ and g are defined above and $\rho = \rho(p + p_0)$.

In order to solve the nonlinear problem (1.16), we shall rephrase it as a fixed-point problem. Let $\mathbf{w}$ be a given vector field with $\mathbf{w} = 0$ on Γ, and let η be a given function with $\eta = 0$ on Γ_{in}. Let

$$\mathbf{F}(\mathbf{w}, \eta) = \mathbf{f} - \rho[(\mathbf{w} \cdot \nabla)\mathbf{w} + (\mathbf{w} \cdot \nabla)\mathbf{u}_0] - \bar{\rho}\eta(\mathbf{u}_0 \cdot \nabla)\mathbf{w}, \tag{1.17}$$

$$G(\mathbf{w}, \eta) = g - \tau\eta \operatorname{div} \mathbf{w} - \mathbf{w} \cdot \nabla p_0, \tag{1.18}$$

where $\rho = \rho(\eta + p_0), \bar{\rho} = \bar{\rho}(\eta + p_0)$, and $\tau = \tau(\eta + p_0)$. Let $\kappa_0 = \kappa(p_0)$ and $\rho_0 = \rho(p_0)$, and let U and V be given functions. Consider the linear problem

$$\begin{cases} -\mu\Delta\mathbf{u} - \nu\nabla \operatorname{div} \mathbf{u} + \rho_0(\mathbf{u}_0 \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{F} \quad \text{in } \Omega, \\ \kappa_0 \operatorname{div} \mathbf{u} + U p_x + V p_y = G \quad \text{in } \Omega, \\ \mathbf{u} = 0 \quad \text{on } \partial\Omega, \\ p = 0 \quad \text{on } \Gamma_{\text{in}}. \end{cases} \tag{1.19}$$

System (1.19) is a linear system of equations which is somewhat more complicated than the Stokes system. The complications are the presence of the convection term in the first equation of (1.19) and, more importantly, the presence of the pressure p terms in the second equation of (1.19). We shall call (1.19) a compressible Stokes system. If this system has a solution for $\mathbf{F} = \mathbf{F}(\mathbf{w}, \eta)$, $G = G(\mathbf{w}, \eta)$, $U = w_1 + u_0$, and $V = w_2$, we may consider the solution as defining a map $(\mathbf{w}, \eta) \longrightarrow (\mathbf{u}, p)$. If it happens that this map has a fixed point $(\mathbf{u}, p)$, then system (1.19) becomes (1.16). Hence to solve (1.16), it suffices to find a fixed point of this map.

Throughout this paper, we will assume that

$$U \geq C_0 > 0 \tag{1.20}$$

for some constant C_0.

In this paper, we will use the following spaces and norms:

$$\begin{aligned}
\|u\|_0 &\equiv \left\{ \int_\Omega |u(\mathrm{x})|^2 \mathrm{dx} \right\}^{1/2} \quad \text{and} \quad \|u\|_{0,q} \equiv \left\{ \int_\Omega |u(\mathrm{x})|^q \mathrm{dx} \right\}^{1/q}, \\
L^2(\Omega) &= \{u : \|u\|_0 < \infty\} \;\; \text{and} \;\; L^q(\Omega) = \{u : \|u\|_{0,q} < \infty\}, \\
\|u\|_k &\equiv \sum_{j=0}^k \|\nabla^j u\|_0, \qquad \|u\|_{k,q} \equiv \sum_{j=0}^k \|\nabla^j u\|_{j,q}, \\
H^k(\Omega) &\equiv \{u \in L^2(\Omega) : \|u\|_k < \infty\}, \\
H_0^k(\Omega) &\equiv \{u \in H^k(\Omega) : u = 0 \text{ on } \partial\Omega\}, \\
H^{k,q}(\Omega) &\equiv \{u \in L^q(\Omega) : \|u\|_{k,q} < \infty\}, \\
H_0^{k,q}(\Omega) &\equiv \{u \in H^{k,q}(\Omega) : u = 0 \text{ on } \partial\Omega\}, \\
\|u\|_{-1} &\equiv \sup\{\langle u, v\rangle : v \in H_0^1(\Omega), \|v\|_1 = 1\}, \\
\|u\|_{-1,q} &\equiv \sup\{\langle u, v\rangle : v \in H_0^{1,q'}(\Omega), \|v\|_{1,q'} = 1, \; 1/q + 1/q' = 1\}, \\
|u|_{0,\infty} &\equiv \sup\{|u(\mathrm{x})| : \mathrm{x} \in \Omega\}, \\
|u|_{k,\infty} &\equiv \sum_{j=0}^k |\nabla^j u|_{0,\infty}.
\end{aligned}$$

In our proofs, C denotes a generic constant depending on certain quantities. We shall make this dependence explicitly—for example, writing $C(\Omega)$ if C depends only on Ω (for example, in the Sobolev inequalities) or $C(\Omega, u_0, C_0)$ if C depends on Ω, u_0, and C_0, and so on.

2. Large viscosity and constant ambient flows. In this section, we prove the theorems stated above. We start with some lemmas. In these lemma, the functions $\delta_-(y)$ and $d(y)$, which were defined in section 1, are regarded as functions on Ω which are independent of x.

LEMMA 2.1. *Let q be given with $2 < q < 3$. Then δ'_-, $d^2\delta''_- \in \text{\L}^q(\Omega)$.*

Proof. Since the behavior of $\delta'_-(y)$ is like that of $(\delta_-(y) - x^*)/(y - y^*)$ near x^* and $d(y) = |\delta_-(y) - x^*|\sqrt{\epsilon'(\xi)^2 + 1}$ for $\bar{x} \le \xi \le x^*$, using Condition B,

$$\int_{\delta_-(y)}^{x^*} |\delta'_-(y)|^q \mathrm{dx} \le C|y - y^*|^{\frac{1-q}{2}} \quad \text{for some constant } C.$$

Integrating both sides of above inequality with respect to y in a neighborhood of y^*, we have $\delta'_-(y) \in L^q(\Omega)$ for $q < 3$. A similar argument holds near y_*. From the relation $y = \epsilon(x), x = \delta_-(y)$, $\delta''_-(y) = -\epsilon''(\delta_-(y))\delta'_-(y)^3$. A similar argument shows that $d(y)^2\delta''_-(y) \in L^q(\Omega)$. □

In order to study the linear system (1.19), we first consider the continuity equation

$$\begin{aligned} p_x(x,y) + U^{-1}V(x,y)p_y(x,y) &= \hat{H}(x,y) \quad \text{in } \Omega, \\ p(\delta_-(y), y) &= 0, \end{aligned} \tag{2.1}$$

where U and V are given and $\hat{H} = U^{-1}H$. To solve the continuity equation, consider for each fixed $\bar{y}$,

$$\begin{aligned} p_x(x, h(x,\bar{y})) + U^{-1}V(x, h(x,\bar{y}))p_y(x, h(x,\bar{y})) &= \hat{H}(x, h(x,\bar{y})) \quad \text{in } \Omega, \\ p(\delta_-(\bar{y}), \bar{y}) &= 0, \end{aligned} \tag{2.2}$$

where $h(x,\bar{y})$ is the solution of the following first-order ODE: for each $\bar{y}$,

$$\begin{aligned} h_x(x,\bar{y}) &= U^{-1}V(x, h(x,\bar{y})) \quad \text{in } \Omega, \\ h(\delta_-(\bar{y}), \bar{y}) &= \bar{y}. \end{aligned} \tag{2.3}$$

The solution $h(x,\bar{y})$ of (2.3) is given by

$$h(x,\bar{y}) = \int_{\delta_-(\bar{y})}^{x} U^{-1}V(s, h(s,\bar{y}))ds + \bar{y}. \tag{2.4}$$

Now we set $y = h(x,\bar{y})$. Then $\bar{y} = \varphi(x,y)$ for some function φ since $h(x,\bar{y})$ is increasing in $\bar{y}$. Thus

$$y = \int_{\delta_-(\varphi(x,y)}^{x} U^{-1}V(s, h(s, \varphi(x,y)))ds + \varphi(x,y). \tag{2.5}$$

Differentiating both sides of (2.5) with respect to x and y, respectively, and using $U^{-1}V = 0$ on Γ, we get

$$\varphi_x(x,y) = -U^{-1}V(x,y) - \int_{\delta_-(\varphi(x,y))}^{x} \frac{\partial}{\partial x}[U^{-1}V(s, h(s, \varphi(x,y)))]ds, \tag{2.6}$$

$$\varphi_y(x,y) = 1 - \int_{\delta_-(\varphi(x,y))}^{x} \frac{\partial}{\partial y}[U^{-1}V(s, h(s, \varphi(x,y)))]ds. \tag{2.7}$$

LEMMA 2.2. *Let h and φ be given as above. Then $\|h\|_{2,q}$, $\|\varphi\|_{2,q}$, and $\|\nabla(\delta_- \circ \varphi)\|_{0,q}$ are bounded by a constant of the form $C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0)$.*

Proof. The proof follows from the formulas above, in particular, Lemma 2.1, (2.6), (2.7), and the fact that $V = 0$ on Γ. □

Now using (2.2) $\sim$ (2.5), the solution of (2.1) is given by

$$p(x,y) = \int_{\delta_-(\varphi(x,y))}^{x} \hat{H}(s, h(s, \varphi(x,y)))ds, \tag{2.8}$$

where $\hat{H} \equiv U^{-1}H$.

LEMMA 2.3. *If $2 < q < 3$, then the solution p of (2.1), given by (2.8), satisfies the following inequality:*

$$\|p\|_{1,q} \leq C\|H\|_{1,q}, \tag{2.9}$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0)$.

Proof. Differentiating both sides of (2.8) with respect to y, we get

$$p_y(x,y) = \int_{\delta_-(\varphi(x,y))}^{x} \frac{\partial}{\partial y}\hat{H}(s, h(s, \varphi(x,y)))ds - \hat{H}(\delta_-(\varphi(x,y)), \varphi(x,y))\frac{\partial \delta_-(\varphi(x,y))}{\partial y},$$

and for simplicity, writing $\varphi(x,y)$ as φ, we have

$$\begin{aligned} |p_y(x,y)|^q &\leq C\left\{\left[\int_{\delta_-(\varphi(x,y))}^{x} \left|\frac{\partial}{\partial y}\hat{H}(s, h(s, \varphi(x,y)))\right| ds\right]^q + |\hat{H}(\delta_-(\varphi), \varphi)|^q \left|\frac{\partial \delta_-(\varphi)}{\partial y}\right|^q\right\} \\ &\leq C\left\{\int_{\delta_-(\varphi(x,y))}^{x} \left|\frac{\partial}{\partial y}\hat{H}(s, h(s, \varphi(x,y)))\right|^q ds + \left|\hat{H}\right|_\infty^q \left|\frac{\partial \delta_-(\varphi)}{\partial y}\right|^q\right\}, \end{aligned} \tag{2.10}$$

where we used $(a+b)^q \leq 2^{q-1}(a^q + b^q)$ in the first inequality and Hölder's inequality in the second inequality and C depends on Ω. Next, integrating both sides of (2.10) with respect to x from $\delta_-(y)$ to $\delta_+(y)$,

$$\begin{aligned} \int_{\delta_-(y)}^{\delta_+(y)} |p_y(x,y)|^q dx &\leq C\int_{\delta_-(y)}^{\delta_+(y)} \int_{\delta_-(\varphi(x,y))}^{x} \left|\frac{\partial}{\partial y}\hat{H}(s, h(s, \varphi(x,y)))\right|^q dsdx (\equiv A(y)) \\ &\quad + C|\hat{H}|_\infty^q \int_{\delta_-(y)}^{\delta_+(y)} \left|\frac{\partial \delta_-(\varphi)}{\partial y}\right|^q dx (\equiv B(y)). \end{aligned} \tag{2.11}$$

Integrating $A(y)$ and changing the variables of integration, we get

$$\begin{aligned} \int_{y_*}^{y^*} A(y)dy &= C\int_{y_*}^{y^*} \int_{\delta_-(y)}^{\delta_+(y)} \left\{\int_{\delta_-(\varphi(x,y))}^{x} \left|\frac{\partial}{\partial y}\hat{H}(s, h(s, \varphi(x,y)))\right|^q ds\right\} dxdy \\ &\leq C(\|U\|_{2,q}, \|V\|_{2,q}, \Omega)\int_\Omega |\hat{H}_y(x,y)|^q \mathrm{dx}. \end{aligned} \tag{2.12}$$

In order to estimate $B(y)$ in (2.11), we observe that $\partial(\delta_- \circ \varphi)/\partial y$ is close to $\delta'_-(y)$ near Γ. Then we have

$$\int_{y=y_*}^{y^*} \int_{\delta_-(y)}^{\delta_+(y)} \left|\frac{\partial \delta_-(\varphi(x,y))}{\partial y}\right|^q dxdy \leq C\|\delta'_-(y)\|_{0,q} < \infty, \tag{2.13}$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0)$ and we have used Lemma 2.1. Since $q > 2$, by the Sobolev embedding theorem, $H^{1,q}(\Omega) \subseteq C^0(\Omega)$,

$$|\hat{H}|_\infty^q \leq C\|\hat{H}\|_{1,q}^q \quad \text{for some constant } C. \tag{2.14}$$

Combining (2.11)–(2.14), we get $\|p_y\|_q^q \leq C\|\hat{H}\|_{1,q}^q$ for some constants C. Similarly, applying the same procedure for p_x, we get $\|p_x\|_q^q \leq C\|\hat{H}\|_{1,q}^q$. Furthermore, we can easily get the L^q-estimate for $p(x, y)$. Thus, since $\hat{H} = U^{-1}H$, the result follows. □

Consider the following elliptic boundary value problem:

$$\begin{aligned} -\mu\Delta\mathbf{u} - \nu\nabla\mathrm{div}\,\mathbf{u} + \rho_0(\mathbf{u}_0 \cdot \nabla)\mathbf{u} &= \bar{\mathbf{F}} \quad \text{in } \Omega, \\ \mathbf{u} &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.15}$$

Remark. The ellipticity of the operator follows from the inequalities satisfied by μ and ν. For simplicity, we set $\bar{\mu} = (\mu + \nu)/\mu$.

Lemma 2.4. *Assume that $|\nabla(\rho_0\mathbf{u}_0)|_\infty$ is small enough. Then there is a unique weak solution $\mathbf{u}$ of* (2.15) *with* (i) $\mu\|\mathbf{u}\|_1 \leq C_1\|\bar{\mathbf{F}}\|_0$, *where $C_1 = C(\bar{\mu}, |\nabla(\rho_0\mathbf{u}_0)|_\infty)$. Furthermore, the solution $\mathbf{u}$ of* (2.15) *satisfies* (ii) $\mu\|\mathbf{u}\|_{k,q} \leq C_2\|\bar{\mathbf{F}}\|_{k-2,q}$, *where $C_2 = C(\Omega, \bar{\mu}, \mathbf{u}_0, \rho_0, k)$, k is an integer ≥ 1, $1 < q < \infty$, C_2 is bounded provided that $\bar{\mu}$ ranges over a compact subset of* $(0, \infty)$, *and μ is bounded away from zero. Also, the solution $\mathbf{u}$ of* (2.15) *satisfies* (ii) *with the small condition of $|\rho_0|_{k-2,\infty}$ $(k \geq 2)$.*

Proof. First, let $\bar{\mathbf{u}} = \mu\mathbf{u}$ and consider the modified equation of (2.15); $-\Delta\bar{\mathbf{u}} - (\bar{\mu} - 1)\nabla\mathrm{div}\bar{\mathbf{u}} + \mu^{-1}\rho_0(\mathbf{u}_0 \cdot \nabla)\bar{\mathbf{u}} = \bar{\mathbf{F}}$ in Ω and $\bar{\mathbf{u}} = 0$ on Γ. Next, we multiply both sides by $\bar{\mathbf{u}}$ and integrate by parts. Using the inequalities satisfied by μ and $\bar{\mu}$ and the smallness assumption on $|\nabla(\rho_0\mathbf{u}_0)|_\infty$ (or the condition that μ is large enough), we easily get (i). The existence of a weak solution $\bar{\mathbf{u}} \in H_0^1$ to (i) then follows from the Lax–Milgram lemma. From the standard L^q theory of elliptic equations (see [2]) applied to the modified version of (2.15), we obtain

$$\|\bar{\mathbf{u}}\|_{k,q} \leq \mu^{-1}C(\bar{\mu})|\rho_0\mathbf{u}_0|_{k-2,\infty}\|\bar{\mathbf{u}}\|_{k-1,q} + C(\bar{\mu})\|\bar{\mathbf{F}}\|_{k-2,q}.$$

This inequality holds uniformly if $\bar{\mu}$ ranges over a compact subset of $(0, \infty)$. Suppose that $|\nabla(\rho_0\mathbf{u}_0)|_\infty$ is small enough so that (i) holds. A standard argument by contradiction then establishes (ii). □

The next theorem gives a solution to the linear system (1.19).

Theorem 2.1. *Suppose that $U \geq C_0$. Assume that $|\nabla(\rho_0\mathbf{u}_0)|_\infty$ is small enough. Assume that $\|\kappa_0\|_{1,q}$ is sufficiently small or μ is large enough. Then there exists a unique solution $(\mathbf{u}, p) \in H_0^{2,q} \times H^{1,q}$ of system* (1.19) *with the following a priori estimate:*

$$\mu\|\mathbf{u}\|_{2,q} + \|p\|_{1,q} \leq C(\|\mathbf{F}\|_{0,q} + \|G\|_{1,q}), \tag{2.16}$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0, \bar{\mu}, \|\kappa_0\|_{1,q}, \mathbf{u}_0, \rho_0)$.

Proof. Let $\mathbf{w} \in H_0^{2,q}$ be given, and consider the following problem:

$$\begin{cases} -\mu\Delta\mathbf{u} - \nu\nabla\mathrm{div}\,\mathbf{u} + \rho_0(\mathbf{u}_0 \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{F} \quad \text{in } \Omega, \\ Up_x + Vp_y = G - \kappa_0\mathrm{div}\,\mathbf{w} \quad \text{in } \Omega, \\ \mathbf{u} = 0 \quad \text{on } \partial\Omega, \\ p = 0 \quad \text{on } \Gamma_{\text{in}}. \end{cases} \tag{2.17}$$

Using Lemma 2.3 with $H = G - \kappa_0\mathrm{div}\,\mathbf{w}$, we find that the solution p in (2.1) exists and satifies

$$\|p\|_{1,q} \leq C(\|\kappa_0\|_{1,q}\|\mathbf{w}\|_{2,q} + \|G\|_{1,q}), \tag{2.18}$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0)$. However, by (2.8) with $H = G - \kappa_0 \operatorname{div} \mathbf{w}$, the solution p of the continuity equation (2.17) is uniquely determined, and using this solution p and Lemma 2.4, the problem

$$\begin{aligned} -\mu\Delta\mathbf{u} - \nu\nabla\operatorname{div}\mathbf{u} + \rho_0(\mathbf{u}_0\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{F} \quad \text{in } \Omega, \\ \mathbf{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.19}$$

has a unique solution $\mathbf{u} \in H^{2,q}(\Omega) \cap H_0^{1,q}(\Omega)$ and satisfies the inequality

$$\mu\|\mathbf{u}\|_{2,q} \leq C(\|\nabla p\|_{0,q} + \|\mathbf{F}\|_{0,q}), \tag{2.20}$$

where $C = C(\Omega, \bar{\mu}, \rho_0, \mathbf{u}_0, q)$. Here we combine (2.18) and (2.20) and get

$$\mu\|\mathbf{u}\|_{2,q} + \|p\|_{1,q} \leq C(\|\kappa_0\|_{1,q}\|\mathbf{w}\|_{2,q} + \|\mathbf{F}\|_{0,q} + \|G\|_{1,q}), \tag{2.21}$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0, \bar{\mu}, \rho_0, \mathbf{u}_0, q)$. At this point, we call attention to the linear map $(\mathbf{F}, G, U, V, \mathbf{w}) \longmapsto (\mathbf{u}, p)$ defined by (2.17). If $(\mathbf{u}^*, p^*)$ is the solution corresponding to the data $(\mathbf{F}, G, U, V, \mathbf{w}^*)$, it follows that $(\mathbf{u}-\mathbf{u}^*, p-p^*)$ is the solution corresponding to the data $(0, 0, 0, 0, \mathbf{w} - \mathbf{w}^*)$. Now using (2.21), we get

$$\mu\|\mathbf{u} - \mathbf{u}^*\|_{2,q} + \|p - p^*\|_{1,q} \leq C\|\kappa_0\|_{1,q}\|\mathbf{w} - \mathbf{w}^*\|_{2,q},$$

where $C = C(\Omega, \|U\|_{2,q}, \|V\|_{2,q}, C_0, \bar{\mu}, |\rho_0|_\infty, |\mathbf{u}_0|_\infty, q)$. Thus we have

$$\mu\|\mathbf{u} - \mathbf{u}^*\|_{2,q} \leq C\|\kappa_0\|_{1,q}\|\mathbf{w} - \mathbf{w}^*\|_{2,q}.$$

Hence for fixed $\mathbf{F}$, G, U, and V, if $\mu =$ large or $\|\kappa_0\|_{1,q} =$ small, then the map $\mathbf{w} \longmapsto \mathbf{u}$ is a contraction in the topology of $H_0^{2,q}$, i.e., $\|\mathbf{u}-\mathbf{u}^*\|_{2,q} \leq (1/2)\|\mathbf{w}-\mathbf{w}^*\|_{2,q}$ and it has a unique fixed point $\mathbf{u} = \mathbf{w}$. Thus using this fixed point $\mathbf{u}$, p is uniquely determined by the continuity equation in (2.17) and (2.8). Hence $(\mathbf{u}, p)$ is the desired solution of (1.19). Again, if μ is large or $\|\kappa_0\|_{1,q}$ is small, we get (2.16) by (2.21). □

We give an example of a solution to (1.19) that emphasizes that the restriction on the regularity of the solution $(\mathbf{u}, p)$ of (1.19) in Theorem 2.1 comes from the geometry of $\partial\Omega$, not the regularity of the data. For example, let Ω be the unit disk and let $\mu = 1$, $\nu = 0$, $\mathbf{u}_0 = [1, 0]$, $\kappa_0 = 1$, $U = 1$, and $V = 0$. Let $[u, v, p]$ be defined by

$$\begin{aligned} u(x,y) &= 1 - x^2 - y^2, \\ v(x,y) &= \chi(y)[(1-y^2)^{3/2} - |x|^3], \\ p(x,y) &= 3\chi(y)(x-1)[(1-y^2)^{1/2} + x]. \end{aligned}$$

Here $\chi(y)$ is a smooth function which is $\equiv 1$ near $y = 1$ and $\equiv 0$ near $y = -1$. It is easily seen that $v \in H^{2,q}(\Omega)$ and $p \in H^{1,q}(\Omega)$ for $q < 3$. Also, $u = v = 0$ on Γ and $p = 0$ on Γ_{in}. For y near 1, $f_1 = -\Delta u + u_x + p_x = 3(1-y^2)^{1/2} + 4x + 1 \in H^{1,q}(\Omega)$ and $f_2 = -\Delta v + v_x + p_y = 3(y-y^2-xy)(1-y^2)^{-1/2} + 3(1-y^2)^{1/2} + 6|x| - 3x|x|$. The first term in this formula for f_2 may be written $3y(1-y)(1-y^2)^{-1/2} - 3xy(1-y^2)^{-1/2} \in H^{1,q}(\Omega)$. It may be seen that $|x| \in H^{1,q}(\Omega)$ and $x|x| \in H^{1,q}(\Omega)$. Hence $f_2 \in H^{1,q}(\Omega)$. Finally, $g = u_x + v_y + p_x = 3(1-y)(1-y^2)^{1/2} + 4x - 3 \in H^{2,q}(\Omega)$. Summarizing the computation above, $f_1 \in H^{1,q}(\Omega), f_2 \in H^{1,q}(\Omega)$, and $g \in H^{2,q}(\Omega)$. Hence (u, v, p) satisfies (1.19) with data of sufficient regularity to permit u, v, and p to have more regularity than they have. The singularity of the solution at (x^*, y^*) blocks further regularity in the solution.

Recall that in (1.12) and (1.13), we defined $\mathbf{f} = \mu\Delta\mathbf{u}_0 + \nu\nabla\mathrm{div}\,\mathbf{u}_0 - \nabla p_0 - \rho(\mathbf{u}_0 \cdot \nabla)\mathbf{u}_0$ and $g = -\kappa(\eta + p_0)\mathrm{div}\,\mathbf{u}_0 - \mathbf{u}_0 \cdot \nabla p_0$. Hence we get

$$\|\mathbf{f}\|_{0,q} + \|g\|_{1,q} \le C(\|\nabla\mathbf{u}_0\|_{1,q} + \|\nabla p_0\|_{1,q}). \tag{2.22}$$

Let us set $M \equiv \|\nabla\mathbf{u}_0\|_{1,q} + \|\nabla p_0\|_{1,q}$. In order to solve system (1.16), we use both the linear system (1.19) with $U = w_1 + u_0$ and $V = w_2$ and (2.16). Before doing this, we need the following lemma.

LEMMA 2.5. *Let* $\mathbf{w} \in H_0^{2,q}$ *and* $\eta \in H^{1,q}$. *Let* $\mathbf{F} = \mathbf{F}(\mathbf{w}, \eta)$ *and* $G = G(\mathbf{w}, \eta)$ *be given by* (1.17) *and* (1.18). *Let* $U = w_1 + u_0$ *and* $V = w_2$. *Then the solution* $(\mathbf{u}, p) \in H_0^{2,q} \times H^{1,q}$ *of system* (1.19) *that is given by Theorem* 2.1 *satisfies the inequality*

$$\|\mathbf{u}\|_{2,q} + \mu^{-1}\|p\|_{1,q} \le C(\|\mathbf{w}\|_{2,q} + \mu^{-1}\|\eta\|_{1,q})^2 + \frac{CM}{\mu}\|\mathbf{w}\|_{1,q} + \frac{CM}{\mu}, \tag{2.23}$$

where $C = C(\Omega, \bar{\mu}, \|\mathbf{w}\|_{2,q}, \|\mathbf{u}_0\|_{2,q}, C_0, |\kappa_0|_{1,\infty}, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$.

Proof. Using formula (1.17) for $\mathbf{F}(\mathbf{w}, \eta)$ and the Sobolev embedding $H^{1,q} \subseteq C^0$,

$$\begin{aligned} \|\mathbf{F}(\mathbf{w}, \eta)\|_{0,q} &\le C(\|\mathbf{f}\|_{0,q} + \|\mathbf{w} \cdot \nabla\mathbf{w}\|_{0,q} + \|\nabla\mathbf{w}\|_{0,q}|\eta|_\infty + \|\mathbf{w} \cdot \nabla\mathbf{u}_0\|_{0,q}) \\ &\le C(\|\mathbf{w}\|_{1,q} + \|\eta\|_{1,q} + \|\nabla\mathbf{u}_0\|_{1,q})\|\mathbf{w}\|_{1,q} + C\|\mathbf{f}\|_{0,q}, \end{aligned} \tag{2.24}$$

where $C = C(\Omega, |\rho|_\infty, |\bar{\rho}|_\infty)$. Next, using (1.18), we estimate $\|G(\mathbf{w}, \eta)\|_{1,q}$ and get

$$\|G\|_{1,q} \le C(\|\tau\eta\mathrm{div}\,\mathbf{w}\|_{1,q} + \|\mathbf{w} \cdot \nabla p_0\|_{1,q} + \|g\|_{1,q}), \tag{2.25}$$

$$\begin{aligned} \|\nabla(\eta\mathrm{div}\,\mathbf{w})\|_{0,q} &\le C(\|\nabla\eta\mathrm{div}\,\mathbf{w}\|_{0,q} + \|\eta\nabla\mathrm{div}\,\mathbf{w}\|_{0,q}) \\ &\le C(|\mathrm{div}\,\mathbf{w}|_\infty\|\nabla\eta\|_{0,q} + |\eta|_\infty\|\mathbf{w}\|_{2,q}), \end{aligned} \tag{2.26}$$

$$\|\eta\mathrm{div}\,\mathbf{w}\|_{0,q} \le |\mathrm{div}\,\mathbf{w}|_\infty\|\eta\|_{0,q}. \tag{2.27}$$

Hence, using the inequality $\|\mathbf{w} \cdot \nabla p_0\| \le C\|\nabla p_0\|_{1,q}\|\mathbf{w}\|_{1,q}$,

$$\|G\|_{1,q} \le C(\|\mathbf{w}\|_{2,q}\|\eta\|_{1,q} + \|\nabla p_0\|_{1,q}\|\mathbf{w}\|_{1,q} + \|g\|_{1,q}). \tag{2.28}$$

Finally, combining (2.24) and (2.28), we have

$$\begin{aligned} \|\mathbf{F}(\mathbf{w}, \eta)\|_{0,q} + \|G(\mathbf{w}, \eta)\|_{1,q} &\le C(\|\mathbf{w}\|_{2,q}\|\eta\|_{1,q} + \|\mathbf{w}\|_{1,q}^2 + M\|\mathbf{w}\|_{1,q}) \\ &\quad + C(\|\mathbf{f}\|_{0,q} + \|g\|_{1,q}), \end{aligned} \tag{2.29}$$

where $C = C(\Omega, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$. Recalling that $U = w_1 + u_0$ and $V = w_2$, recalling $\mathbf{f}$ and g from (1.12) and (1.13), combining (2.16) and (2.29), then dividing the resulting inequality by μ, and finally using $2\alpha\beta \le (\alpha+\beta)^2$, we get the required inequality. □

We now construct a fixed-point map. To do this, we let $\sigma = \mu^{-1}p$ and $\pi = \mu^{-1}\eta$. Then the linear sytem (1.19) becomes

$$\begin{cases} -\Delta\mathbf{u} - \mu^{-1}\nu\nabla\mathrm{div}\,\mathbf{u} + \mu^{-1}\rho_0(\mathbf{u}_0 \cdot \nabla)\mathbf{u} + \nabla\sigma = \mu^{-1}\mathbf{F}(\mathbf{w}, \mu\pi) \quad \text{in } \Omega, \\ \kappa_0\mathrm{div}\,\mathbf{u} + \mu(U\sigma_x + V\sigma_y) = G(\mathbf{w}, \mu\pi) \quad \text{in } \Omega, \\ \mathbf{u} = 0 \quad \text{on } \partial\Omega, \\ \sigma = 0 \quad \text{on } \Gamma_{\mathrm{in}}, \end{cases} \tag{2.30}$$

where $U = w_1 + u_0$ and $V = w_2$.

The existence of the solution $(\mathbf{u}, \sigma)$ of (2.30) is guaranteed by Theorem 2.1. For fixed $\mathbf{f}$ and g, define the map $T : (\mathbf{w}, \pi) \longrightarrow (\mathbf{F}(\mathbf{w}, \mu\pi), G(\mathbf{w}, \mu\pi)) \longrightarrow (\mathbf{u}, \sigma)$, where $(\mathbf{u}, \sigma)$ is the solution of (2.30). We want to prove that the map T is a contraction in some ball B in the topology of $H_0^{1,q} \times H^{0,q}$ if μ is sufficiently large or $\|\kappa_0\|_{1,q}$ is small enough. Choose a ball

$$B \equiv \{(\mathbf{w}, \pi) \in H_0^{2,q} \times H^{1,q} : \|\mathbf{w}\|_{2,q} + \|\pi\|_{1,q} \le A\}, \tag{2.31}$$

where $A \le 1$ will be chosen later.

Lemma 2.6. *Let the constants M and C be given in* (2.22) *and* (2.23)*, respectively. Assume that $\mu^{-1}M \le A/3C$ and $A \le 1/3C$. Then we have $T(B) \subset B$.*

Proof. Let $(\mathbf{w}, \pi) \in B$. Then $\|\mathbf{w}\|_{2,q} + \|\pi\|_{1,q} \le A$. Now from (2.23), we get

$$\begin{aligned}\|\mathbf{u}\|_{2,q} + \|\sigma\|_{1,q} &\le CA^2 + \frac{CM}{\mu}A + \frac{A}{3}\\ &\le \frac{A}{3} + \frac{A^2}{3} + \frac{A}{3}\\ &\le A.\end{aligned}$$

Thus the result follows. □

Remark. If A is small, and if μ is large enough or M is small enough, then we get Lemma 2.6. Thus Lemma 2.6 covers both cases.

Lemma 2.7. *Assume that M and $\|\kappa_0\|_{1,q}$ are small enough or μ is large enough. Then for fixed $\mathbf{f}$ and g, the map $T : B \longrightarrow B$ is a contraction in the topology of $H_0^{1,q} \times H^{0,q}$ if A is small.*

Proof. Consider $(\mathbf{u}, \sigma) = T(\mathbf{w}, \pi)$ and $(\mathbf{u}^*, \sigma^*) = T(\mathbf{w}^*, \pi^*)$ and set $\mathbf{F} = \mathbf{F}(\mathbf{w}, \mu\pi)$, $\mathbf{F}^* = \mathbf{F}(\mathbf{w}^*, \mu\pi^*)$, $G = G(\mathbf{w}, \mu\pi)$, and $G^* = G(\mathbf{w}^*, \mu\pi^*)$. Then from (2.30), we get

$$\begin{cases} -\Delta(\mathbf{u}-\mathbf{u}^*) - \mu^{-1}\nu\nabla\mathrm{div}(\mathbf{u}-\mathbf{u}^*) + \mu^{-1}\rho_0(\mathbf{u}_0\cdot\nabla)(\mathbf{u}-\mathbf{u}^*)\\ \qquad\qquad + \nabla(\sigma-\sigma^*) = \mu^{-1}(\mathbf{F}-\mathbf{F}^*) \quad \text{in } \Omega,\\ (\mathbf{w}^*+\mathbf{u}_0)\cdot\nabla(\sigma-\sigma^*)\\ \qquad = \mu^{-1}[G-G^*-\kappa_0\mathrm{div}(\mathbf{u}-\mathbf{u}^*)] - \mu^{-1}(\mathbf{w}-\mathbf{w}^*)\cdot\nabla p \quad \text{in } \Omega,\\ \mathbf{u}-\mathbf{u}^* = 0 \quad \text{on } \partial\Omega,\\ \sigma-\sigma^* = 0 \quad \text{on } \Gamma_{\text{in}}. \end{cases} \tag{2.32}$$

Let h be the solution of

$$\begin{aligned} h_x(x,\bar{y}) &= U^{*-1}V^*(x, h(x,\bar{y})),\\ h(\delta_-(\bar{y}),\bar{y}) &= \bar{y}. \end{aligned} \tag{2.33}$$

Applying the same procedures as in (2.3), (2.4), (2.5), and (2.8) to the continuity equation above, we can get

$$(\sigma-\sigma^*)(x,y) = \int_{\delta_-(\varphi(x,y))}^{x} B(s, h(s,\varphi(x,y)))ds, \tag{2.34}$$

where $B \equiv \mu^{-1}U^{*-1}[(G-G^*) - \kappa_0\mathrm{div}(\mathbf{u}-\mathbf{u}^*) - (\mathbf{w}-\mathbf{w}^*)\cdot\nabla p]$, $U^* = w_1^* + u_0$. However,

$$\begin{aligned}\mu^{-1}(\mathbf{F}-\mathbf{F}^*) = &-\rho'(\xi)[(\mathbf{w}\cdot\nabla)\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{u}_0](\pi-\pi^*)\\ &+ \rho\mu^{-1}\{[(\mathbf{w}-\mathbf{w}^*)\cdot\nabla]\mathbf{w} + (\mathbf{w}^*\cdot\nabla)(\mathbf{w}-\mathbf{w}^*) + [(\mathbf{w}-\mathbf{w}^*)\cdot\nabla]\mathbf{u}_0\}\\ &- [\bar{\rho}'(\xi)\eta(\mathbf{u}_0\cdot\nabla)\mathbf{w} + \bar{\rho}(\mathbf{u}_0\cdot\nabla)\mathbf{w}](\pi-\pi^*) - \bar{\rho}\pi^*(\mathbf{u}_0\cdot\nabla)(\mathbf{w}-\mathbf{w}^*),\end{aligned}$$

where we used the mean-value property for ρ and $\bar{\rho}$ and where ξ is between π and π^*. Hence

$$\begin{aligned}\|\mu^{-1}(\mathbf{F}-\mathbf{F}^*)\|_{0,q} &\leq C(\|\mathbf{w}\|_{2,q}^2+\|\mathbf{w}\|_{2,q}\|\eta\|_{1,q}\\ &\quad+\|\mathbf{w}\|_{2,q}+|\nabla\mathbf{u}_0|_{0,\infty}\|\mathbf{w}\|_{2,q})\|\pi-\pi^*\|_{0,q}\\ &\quad+C(\|\mathbf{w}\|_{2,q}+\|\mathbf{w}^*\|_{1,q}+\|\pi^*\|_{1,q}+\mu^{-1}|\nabla\mathbf{u}_0|_{\infty})\|\mathbf{w}-\mathbf{w}^*\|_{1,q}\\ &\leq CA(\|\mathbf{w}-\mathbf{w}^*\|_{1,q}+\|\pi-\pi^*\|_{0,q}) \quad \text{if } \mu^{-1}M\leq A, \end{aligned} \tag{2.35}$$

where $C = C(A,\mu,|\rho'|_{\infty},|\bar{\rho}|_{1,\infty})$ and $M = \|\nabla\mathbf{u}_0\|_{1,q}+\|\nabla p_0\|_{1,q}$. Next, setting $\tau^* = \tau(\eta^*+p_0)$,

$$\begin{aligned}\mu^{-1}(G-G^*) &= -\mu^{-1}[\tau\eta\mathrm{div}\,\mathbf{w}-\tau^*\eta^*\mathrm{div}\mathbf{w}^*+(\mathbf{w}-\mathbf{w}^*)\cdot\nabla p_0]\\ &= -[\tau(\eta+p_0)+\tau'(\xi)\eta^*](\pi-\pi^*)\mathrm{div}\,\mathbf{w}-[\tau\pi^*\mathrm{div}(\mathbf{w}-\mathbf{w}^*)\\ &\qquad+\mu^{-1}(\mathbf{w}-\mathbf{w}^*)\cdot\nabla p_0]. \end{aligned} \tag{2.36}$$

Now

$$\begin{aligned}\|\sigma-\sigma^*\|_{0,q} &\leq C\|B\|_{0,q}\\ &\leq C\left(\left\|\frac{G-G^*}{\mu}\right\|_{0,q}+\frac{|\kappa_0|_{\infty}}{\mu}\|\mathrm{div}(\mathbf{u}-\mathbf{u}^*)\|_{0,q}+\mu^{-1}|\mathbf{w}-\mathbf{w}^*|_{\infty}\|\nabla p\|_{0,q}\right)\\ &\leq C\left(\left\|\frac{G-G^*}{\mu}\right\|_{0,q}+\frac{\|\kappa_0\|_{1,q}}{\mu}\|\mathbf{u}-\mathbf{u}^*\|_{1,q}+\mu^{-1}\|\mathbf{w}-\mathbf{w}^*\|_{1,q}\|\nabla p\|_{0,q}\right), \end{aligned}$$

where we used the Sobolev embedding $H^{1,q}\subseteq C^0$ and $C = C(\Omega, C_0, A)$. From Lemma 2.6, we get $\mu^{-1}\|\nabla p\|_{0,q}\leq A$, and using (2.36),

$$\|\mu^{-1}(G-G^*)\|_{0,q}\leq C|\mathrm{div}\,\mathbf{w}|_{\infty}\|\pi-\pi^*\|_{0,q}+C(\|\pi^*\|_{1,q}+\mu^{-1}\|\nabla p_0\|_{1,q})\|\mathbf{w}-\mathbf{w}^*\|_{1,q},$$

where $C = C(\mu,|\tau|_{1,\infty},A,|\kappa|_{\infty})$. Thus if $\mu^{-1}M\leq A$ and $|\mathrm{div}\,\mathbf{w}|_{\infty}\leq CA$,

$$\begin{aligned}\|\sigma-\sigma^*\|_{0,q} &\leq C\mu^{-1}\|\kappa_0\|_{1,q}\|\mathbf{u}-\mathbf{u}^*\|_{1,q}\\ &\quad+CA(\|\mathbf{w}-\mathbf{w}^*\|_{1,q}+\|\pi-\pi^*|_{0,q}), \end{aligned} \tag{2.37}$$

where $C = C(\Omega, C_0, A, |\tau|_{1,\infty}, |\kappa|_{\infty})$. From the momentum equation, standard elliptic theory, Lemma 2.4, (2.35), (2.37), and the assumption that μ is large or $\|\kappa_0\|_{1,q}$ is small enough, we get

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}^*\|_{1,q} &\leq C(\|\nabla(\sigma-\sigma^*)\|_{-1,q}+\|\mu^{-1}(\mathbf{F}-\mathbf{F}^*)\|_{0,q})\\ &\leq C(\|\sigma-\sigma^*\|_{0,q}+\|\mu^{-1}(\mathbf{F}-\mathbf{F}^*)\|_{0,q})\\ &\leq CA(\|\mathbf{w}-\mathbf{w}^*\|_{1,q}+\|\pi-\pi^*\|_{0,q}), \end{aligned} \tag{2.38}$$

where $C = C(\Omega, C_0, A, |\tau|_{1,\infty}, |\kappa|_{\infty})$. Combining (2.37) and (2.38), we get

$$\|\mathbf{u}-\mathbf{u}^*\|_{1,q}+\|\sigma-\sigma^*\|_{0,q}\leq CA(\|\mathbf{w}-\mathbf{w}^*\|_{1,q}+\|\pi-\pi^*|_{0,q}), \tag{2.39}$$

where C depends on the same quantity as in (2.38). Hence if $A\leq 1/2C$, we get the inequality

$$\|\mathbf{u}-\mathbf{u}^*\|_{1,q}+\|\sigma-\sigma^*\|_{0,q}\leq \frac{1}{2}(\|\mathbf{w}-\mathbf{w}^*\|_{1,q}+\|\sigma-\sigma^*\|_{0,q}). \tag{2.40}$$

This means that $T : B \longrightarrow B$, defined by $T(\mathbf{w}, \pi) = (\mathbf{u}, \sigma)$, is a contraction in the topology of $H^{1,q} \times H^{0,q}$. $\square$

Recall that $\mathbf{u} = \bar{\mathbf{u}} = \mathbf{u} - \mathbf{u}_0$ and $p = \bar{p} = p - p_0$. Combining all of the previous lemma series and theorems, we obtain a result for the existence and regularity of the solution $(\mathbf{u}, p)$ for problem (1.1) in the case of large viscosity.

THEOREM 2.2. *Suppose that* $\mathbf{u}_0 \in H^{2,q}(\Omega)$, $p_0 \in H^{1,q}(\Omega)$. *Then there is a constant* μ^* *depending on* $\|\mathbf{u}_0\|_{2,q} + \|p_0\|_{1,q}$ *such that if* $\mu \geq \mu^*$, *then problem* (1.16) *has a unique solution* $(u, v, p) \in H^{2,q} \times H^{2,q} \times H^{1,q}$ *with the following estimate:*

$$\|u - u_0\|_{2,q} + \|v\|_{2,q} + \|\sigma - \sigma_0\|_{1,q} \leq C_1, \tag{2.41}$$

where $C_1 = C(\Omega, \mu, \nu, \mu^*, C_0, |\kappa_0|_{1,\infty}, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$, $\sigma = \mu^{-1}p$, *and* $\sigma_0 = \mu^{-1}p_0$.

Proof. The ball B defined by (2.31) is a compact convex subset of $H_0^{1,q} \times H^{0,q}$. From Lemma 2.6, $T(B) \subset B$. From Lemma 2.7, the map T is continuous from B to $H_0^{1,q} \times H^{0,q}$. Since the continuous image of a compact set is compact, $T(B)$ is compact in $H_0^{1,q} \times H^{0,q}$. Hence by the Schauder fixed-point theorem, there is a unique $(\mathbf{u}, \sigma) \in B$ with $T(\mathbf{u}, \sigma) = (\mathbf{u}, \sigma)$. Now we combine the previous lemma series and theorems to get our main result. Let $(\mathbf{w}, \pi) \in B$, defined by (2.31). Then $\|\mathbf{w}\|_{2,q} + \|\pi\|_{1,q} \leq A$ and the constant C in (2.23) becomes $C = C(\Omega, \mu, \nu, A, \|\mathbf{u}_0\|_{2,q}, C_0, |\kappa_0|_{1,\infty}, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$. Let us denote this constant by C_3. However, the map T has a unique fixed point $(\mathbf{u}, \sigma) \in B$ with $\mathbf{u} = \mathbf{w}$ and $\sigma = \pi$. Therefore, if $\mathbf{u}_0 \in H^{2,q}$, $p_0 \in H^{1,q}$, $3C_3M/A \leq \mu^* \leq \mu$, and $\mu^{-1}C_3M \leq 1/3$, then using (2.23), the solution $(\mathbf{u}, \sigma)$ of system (1.16) satisfies the following inequality:

$$\|\mathbf{u}\|_{2,q} + \|\sigma\|_{1,q} \leq C_1, \tag{2.42}$$

where $C_1 = C(\Omega, \mu, \nu, \mu^*, C_0, |\kappa_0|_{1,\infty}, |\tau|_{1,\infty}, |\rho|_\infty, |\bar{\rho}|_\infty)$. $\square$

In Theorem 2.2, we have found the solution $(\mathbf{u}, p)$ of the nonlinear system (1.16) with the condition $\mu =$ large. This means that the Reynolds number is small and so the flows in this domain are laminar flows. The flows have a small perturbation around the ambient flows.

Next, combining the lemmas and theorems above, we obtain a result for the existence and regularity of the solution $(\mathbf{u}, p)$ of problem (1.1) in the case of nearly constant ambient flow.

THEOREM 2.3. *Let* $\mathbf{u}_0 \in H^{2,q}(\Omega)$, $p_0 \in H^{1,q}(\Omega)$, *and* $2 < q < 3$. *For any constant* C_1, *there is a constant* C_2 *such that if* $\|\mathbf{u}_0\|_{2,q} + \|p_0\|_{1,q} \leq C_1$ *and* $\|\nabla \mathbf{u}_0\|_{1,q} + \|\nabla p_0\|_{1,q} \leq C_2$, *then there is a unique solution* $(u, v, p) \in H_0^{2,q} \times H_0^{2,q} \times H^{1,q}$ *of system* (1.1) *with the following estimate:*

$$\|u - u_0\|_{2,q} + \|v\|_{2,q} + \|p - p_0\|_{1,q} \leq C_3, \tag{2.43}$$

where $C_3 = C(\Omega, \mu, \nu, C_0, C_1, C_2, |\rho|_\infty, |\bar{\rho}|_\infty, |\tau|_{1,\infty})$.

Proof. The existence of the solution (u, v, p) of (1.1) easily follows from the similar procedures in the proof of Theorem 2.2. Furthermore, inequality (2.43) can be obtained by using the same methods in the proof of Theorem 2.2. $\square$

REFERENCES

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

[2] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying genneral boundary conditions* I, II, Comm. Pure Appl. Math., 12 (1959), pp. 623–727, 17 (1964), pp. 35–92.

[3] H. Beirão da Veiga, *An L^p-theory for the n-dimensional, stationary, compressible Navier–Stokes equations, and the incompressible limit for compressible fluids: The equilibrium solutions*, Comm. Math. Phys., 109 (1987), pp. 229–248.

[4] R. B. Kellogg, *Discontinous solution of the linearized, steady state, compressible viscous, Navier–Stokes equations*, SIAM J. Math. Anal., 19 (1988), pp. 567–579.

[5] A. Valli, *On the existence of stationary solution to compressible Navier–Stokes equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 99–113.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.

[7] J. B. Conway, *A Course in Functional Analysis*, Graduate Texts in Mathematics 96, Springer-Verlag, Berlin, New York, Heidelberg, Tokyo, 1984.

[8] B. Liu and R. B. Kellogg, *Discontinous solution of linearized, steady state, viscous, compressible flows*, J. Math. Anal. Appl., 180 (1993), pp. 469–487.

[9] J. D. Anderson, Jr., *Fundamental of Aerodynamics*, McGraw–Hill, New York, 1984.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 109–135, January 1997

THE RIEMANN PROBLEM FOR AN INHOMOGENEOUS CONSERVATION LAW WITHOUT CONVEXITY*

CARLO SINESTRARI†

Abstract. The paper studies the Riemann problem for a conservation law with a source term and a nonconvex flux-function. The complete solution is provided in the case when the flux has one inflection point and the Riemann states are stationary states of the source term. For small times, the structure of the solutions is similar to the homogeneous case. As the time increases, the size of the shocks may decrease under the action of the source, while rarefaction waves tend to traveling waves. It is also proved that if the flux has more than one inflection point, there may be shocks vanishing in finite time, in contrast to the case when the flux is convex.

Key words. hyperbolic conservation laws, source term, Riemann problem, entropy solutions, extinction of shocks

AMS subject classifications. 35L65, 35B40, 35L67

PII. S003614109427446X

1. Introduction.

Equations of the form

$$\partial_t u(x,t) + \partial_x f(u(x,t)) = g(x,u(x,t)), \quad x \in \mathbb{R}, \quad t \geq 0, \tag{1.1}$$

are usually called *conservation laws with source* or *inhomogeneous conservation laws* and arise in various physical applications. For example, the case $g(x,u) = c(x)h(u)$ with $h, h' \neq 0$ and c with compact support was considered by Liu and Li [13], [12] as a model for transonic nozzle flow. Also, the equation

$$\partial_t u + h(u)\partial_x u = h_0 - h(u),$$

with h a cubic polinomial and h_0 a constant, was studied by some authors (e.g., Knight and Peterson [9], Murray [17], and Bonilla [2]) as a model for the so-called Gunn effect in semiconductor physics. The interesting feature of the behavior of the solutions in these cases is the appearance of new types of asymptotic states that are different from the shock, rarefaction, and N-waves which are familiar from the case where $g = 0$ (see, e.g., [11]). For instance, in the study of nozzle flow, there appears a new family of stationary waves connecting two states at $x = \pm\infty$.

In a different direction, several researchers, including Dafermos [4], Natalini and Tesei [18], Lyberopoulos [14], [16], Fan and Hale [6], [7], and the author [20], [21], have investigated the properties of (1.1) as a general equation without referring to a specific physical application, the goal being to catalog, at least for the case where $g = g(u)$, the possible structure and asymptotic profiles of solutions. All these papers concern the case of a convex f.

In contrast, the case of a nonconvex f has been considered only in the above-mentioned papers [12], [9], [17], and [2], which deal with particular models, and in the study by Lyberopoulos [15] of periodic solutions in the presence of a linear source. Thus we are still far from having a complete description of the general case. This is

* Received by the editors September 19, 1994; accepted for publication (in revised form) September 19, 1995. This research was partially supported by the Italian National Project MURST ("Problemi nonlineari...").

http://www.siam.org/journals/sima/28-1/27446.html

† Dipartimento di Matematica, Università di Roma "Tor Vergata," Via della Ricerca Scientifica, 00133 Rome, Italy (sinestrari@mat.utovrm.it).

because the geometric structure of the solutions when f changes convexity is much more complicated due to the presence of contact discontinuities, and there is a larger variety of asymptotic states.

In this paper, we intend to provide another step toward the understanding of the properties of the solutions of (1.1) with a general nonconvex f and a nonlinear source term g depending on u alone. The analysis is focused on Riemann initial data

$$u(x,0) = \begin{cases} u_l, & x < 0, \\ \\ u_r, & x > 0. \end{cases} \tag{1.2}$$

The importance of Riemann problems is well known in the study of homogeneous conservation laws, $g = 0$. They admit self-similar solutions which represent the time asymptotic state of any reasonable data asymptotic to u_l and u_r at $x = \mp\infty$. When $g \neq 0$, there are also many reasons to study Riemann problems. First, although the solutions are no longer self-similar, it is still possible to compute them explicitly, and this shows how the source term affects the evolution of discontinuities. Moreover, we still expect a study of the asymptotic states of the Riemann data to provide useful information about the large-time behavior of more general solutions.

In the main part of the paper, we restrict ourselves to the case when u_l and u_r are consecutive simple zeros of g and f has exactly one inflection point between u_l and u_r. Of course, the case when u_l and u_r are zeros of g is the most relevant for the study of the asymptotic behavior because for general Riemann data, the left and right state of the solution converge to stationary states of the source term. We explicitly construct the solution of problem (1.1)–(1.2). Our techniques are partly similar to those used by Ballou in [1] to find solutions to homogeneous conservation laws with piecewise-constant initial data. The results can be roughly described as follows. The solution is piecewise smooth, and for small times, it is close to the solution of the homogeneous problem. As time increases, the amplitude of the shock waves may decrease under the action of the source term, and in some cases, it tends to zero. On the other hand, rarefaction waves evolve and converge to traveling waves. Thus the asymptotic profile of the solution is given by a traveling wave, in some cases discontinuous, and can be found by solving a first-order ordinary differential equation. The statements about the asymptotic behavior are given in section 2, while the construction of the solution is done in sections 4 and 5.

Cases when f has a finite number of inflection points and g has a finite number of simple zeros could be handled by the same techniques; the results would be similar and we would again find an asymptotic state given by a superposition of shock waves and traveling waves. However, there is an interesting behavior arising when f has more than one inflection point, namely the extinction of a shock wave in finite time. Such a behavior is analyzed in the last section of the paper. It is shown that the solution of (1.1)–(1.2) may be discontinuous up to a certain time $T^* > 0$ and then become continuous. Such a property was never observed before for solutions of equation (1.1) and is the opposite of what one usually expects for this type of equation, namely the breakdown of classical solutions and the appearance of discontinuities. In this case, a crucial role is played by the source as well as the nonconvexity of f. In fact, if f is convex (see [3]) or if f has one inflection point and $g \equiv 0$ (see [5]), it has been proved that shock waves cannot vanish in finite time, and so the solution can no longer be regular after the formation of singularities.

The results above suggest what the behavior of solutions would be in more general cases. Assume, for instance, that the initial value u_0 is equal to u_l and u_r for $x < -L$

and $x > L$, respectively, and that it is equal to some function of bounded variation in $[-L, L]$. Then we still expect the solution to converge to a superposition of possibly discontinuous traveling waves. This supposition is supported by the results in the convex case (see [21]), where the solutions corresponding to initial data with compact support exhibit a behavior of this kind. Also, a similar property has been obtained in [12] and [13] in the case of nozzle flow. There the asymptotic states are given by a superposition of rarefaction, shock, and stationary waves; this is because the source term has compact support with respect to the x variable and thus the solutions behave as in the homogeneous case outside a compact interval.

2. Statement of the main results. Let us set $u_m := \min\{u_l, u_r\}$, $u_M := \max\{u_l, u_r\}$. In sections 4 and 5, we find the solution of the Cauchy problem (1.1)–(1.2) under the following assumptions.

(H1) $f \in C^2([u_m, u_M])$ and $g \in C^1([u_m, u_M])$.

(H2) There exists $\bar{u} \in \,]u_m, u_M[$ such that

$$f''(u) < 0 \quad \text{for } u \in \,]u_m, \bar{u}[\,,$$

$$f''(u) > 0 \quad \text{for } u \in \,]\bar{u}, u_M[\,.$$

(H3) $g(u_m) = g(u_M) = 0$, $g(u) > 0$ for $u \in \,]u_m, u_M[$, $g'(u_m) \neq 0$, and $g'(u_M) \neq 0$.

The case when f is convex in $[u_m, \bar{u}]$ and concave in $[\bar{u}, u_M]$ can be reduced to our case by taking $-x$ instead of x as spatial coordinate. Similarly, if g is negative in $]u_m, u_M[$, we can take $-u$ as unknown function. In order to exclude trivial cases, we also assume the following.

(H4) The shock wave connecting u_l and u_r does not satisfy the entropy admissibility criterion (see [19]), i.e.,

$$\frac{f(u_l) - f(u_r)}{u_l - u_r} < f'(u_r).$$

The full expression of the solution is lengthy, but its asymptotic state can be found in a simple way. In fact, it turns out that the solution is asymptotic to a possibly discontinuous traveling-wave solution of (1.1). We first give the result in the case where $u_l < u_r$.

Theorem 2.1. *Assume that hypotheses* (H1)–(H4) *hold and that $u_l < u_r$. Let $\hat{v}$ be the unique value in $]\bar{u}, u_r[$ such that*

$$f'(\hat{v}) = \frac{f(u_l) - f(\hat{v})}{u_l - \hat{v}}$$

and let $\phi : [0, \infty[\,\to [\hat{v}, u_r[$ be defined by

$$\int_{\hat{v}}^{\phi(\xi)} \frac{f'(v) - f'(\hat{v})}{g(v)}\, dv = \xi \quad \textit{for } \xi \geq 0.$$

Then the solution of problem (1.1)–(1.2) *satisfies*

$$u(x,t) = \begin{cases} u_l, & x < f'(\hat{v})t, \\ \phi(x - f'(\hat{v})t) + o(1), & x > f'(\hat{v})t. \end{cases}$$

Here and in the following, we denote by $o(1)$ an error term converging to zero as $t \to \infty$ uniformly in x. Observe that ϕ solves the ordinary differential problem

$$\begin{cases} \phi'(\xi) = \dfrac{g(\phi(\xi))}{f'(\phi(\xi)) - f'(\hat{v})}, & \xi > 0, \\ \lim\limits_{\xi \to 0^+} \phi(\xi) = \hat{v}, \end{cases}$$

and so $\phi(x - f'(\hat{v})t)$ is a traveling-wave solution of equation (1.1).

For the reader's convenience, we recall the results in the homogeneous case (see [8]).

THEOREM 2.2. *Assume that hypotheses* (H1), (H2), *and* (H4) *hold, that* $g \equiv 0$, *and that* $u_l < u_r$. *Then the solution of problem* (1.1)–(1.2) *is*

$$u(x,t) = \begin{cases} u_l, & x < f'(\hat{v})t, \\ h\left(\dfrac{x}{t}\right), & f'(\hat{v})t < x \leq f'(u_r)t, \\ u_r, & x > f'(u_r)t, \end{cases}$$

where $\hat{v}$ *is defined as in Theorem* 2.1 *and* $h : [f'(\hat{v}), f'(u_r)] \to [\hat{v}, u_r]$ *is the inverse of* f'.

Thus in the homogeneous case, the solution is given by a shock wave connecting u_l and $\hat{v}$ and a rarefaction wave connecting $\hat{v}$ and v_r. In the presence of the source term, the rarefaction wave asymptotically becomes a traveling wave, while the size of the shock wave remains unchanged.

Let us now turn to the case $u_l > u_r$.

THEOREM 2.3. *Assume that hypotheses* (H1)–(H4) *hold and that* $u_l > u_r$. *Let* u *be the solution of problem* (1.1)–(1.2).

(i) *Suppose* $f'(u_l) = f'(u_r)$. *Let*

$$L = \int_{u_r}^{u_l} \frac{f'(u_r) - f'(v)}{g(v)} \, dv$$

and let $\psi : [-L, 0] \to [u_r, u_l]$ *be defined by*

$$\int_{u_r}^{\psi(\xi)} \frac{f'(v) - f'(u_r)}{g(v)} \, dv = \xi \tag{2.1}$$

for $\xi \in [-L, 0]$. *Then*

$$u(x,t) = \begin{cases} u_l, & x < f'(u_r)t - L, \\ \psi(x - f'(u_r)t) + o(1), & f'(u_r)t - L \leq x \leq f'(u_r)t, \\ u_r, & x > f'(u_r)t. \end{cases}$$

(ii) *Suppose* $f'(u_l) > f'(u_r)$. *Denote by* $\bar{w}$ *the unique value in* $]\bar{u}, u_l[$ *which satisfies*

$$\frac{f(u_l) - f(\bar{w})}{u_l - \bar{w}} = f'(u_r)$$

and set

$$L = \int_{u_r}^{\bar{w}} \frac{f'(u_r) - f'(v)}{g(v)}\, dv.$$

Let $\psi : [-L, 0] \to [u_r, \bar{w}]$ *be defined by* (2.1) *for all* $\xi \in [-L, 0]$. *Then there exists a function* $\gamma : [0, \infty[\ \to \mathrm{I\!R}$ *such that*

$$\lim_{t\to\infty} \gamma(t) - f'(u_r)t = -L, \qquad \lim_{t\to\infty} \gamma'(t) = f'(u_r),$$

$$u(x,t) = \begin{cases} u_l, & x < \gamma(t), \\ \psi(x - f'(u_r)t) + o(1), & \gamma(t) < x \le f'(u_r)t, \\ u_r, & x > f'(u_r)t. \end{cases}$$

(iii) *Suppose* $f'(u_l) < f'(u_r)$. *Let* $\psi :]-\infty, 0] \to [u_r, u_l[$ *be defined by equality* (2.1) *for all* $\xi \le 0$. *Then*

$$u(x,t) = \begin{cases} \psi(x - f'(u_r)t) + o(1), & x \le f'(u_r)t, \\ u_r, & x > f'(u_r)t. \end{cases}$$

In the homogeneous case, there is the following result (see [8]).

THEOREM 2.4. *Suppose that assumptions* (H1), (H2), *and* (H4) *hold, that* $g \equiv 0$, *and that* $u_l > u_r$. *Then the solution of problem* (1.1)–(1.2) *is*

$$u(x,t) = \begin{cases} u_l, & x < f'(\hat{w})t, \\ j\left(\dfrac{x}{t}\right), & f'(\hat{w})t < x \le f'(u_r)t, \\ u_r, & x > f'(u_r)t, \end{cases}$$

where $\hat{w} \in\ ]u_r, \hat{u}[$ *is the unique value satisfying*

$$f'(\hat{w}) = \frac{f(u_l) - f(\hat{w})}{u_l - \hat{w}}$$

and $j : [f'(\hat{w}), f'(u_r)] \to [u_r, \hat{w}]$ *is the inverse of* f'.

If we look back at Theorem 2.3, we do not find in the asymptotic profile of the solution the shock wave $x = f'(\hat{w})t$ connecting the states u_l and $\hat{w}$, which appears in the homogeneous case. In section 5, we will see that if $g \neq 0$, there is still a shock wave connecting initially the states u_l and $\hat{w}$, but its amplitude decreases with time. The left state remains u_l, but the right state increases monotonically and tends to $\bar{w}$ in case (ii) and to u_l in cases (i) and (iii). Thus in cases (i) and (iii), the amplitude of the shock tends to zero and the asymptotic profile of the solution is continuous. Like the function ϕ of the case where $u_l > u_r$, the function ψ of Theorem 2.3 solves the differential equation which yields the traveling-wave solutions of (1.1).

In section 6, we consider a case where f has two inflection points, and we show that the solution has a shock which vanishes in finite time, as mentioned in section 1. The result that we obtain is the following.

Theorem 2.5. *Let assumptions* (H1) *and* (H3) *hold. Suppose, in addition, that* $u_l < u_r$ *and that* f *satisfies the following properties.*

(H5) *There exist* $\bar{u}_1$ *and* $\bar{u}_2$ *such that*

$$f''(u) > 0 \quad \textit{for } u \in \,]u_l, \bar{u}_1[\,,$$

$$f''(u) < 0 \quad \textit{for } u \in \,]\bar{u}_1, \bar{u}_2[\,,$$

$$f''(u) > 0 \quad \textit{for } u \in \,]\bar{u}_2, u_r[\,.$$

(H6) *There exist* $\hat{v}_1 \in \,]u_l, \bar{u}_1[$ *and* $\hat{v}_2 \in \,]\bar{u}_2, u_r[$ *such that*

$$f'(\hat{v}_1) = f'(\hat{v}_2) = \frac{f(\hat{v}_2) - f(\hat{v}_1)}{\hat{v}_2 - \hat{v}_1}.$$

(H7) $f'(u_l) < f'(\bar{u}_2)$.

Then the solution of problem (1.1)–(1.2) *has the following property: there exists* $T^* > 0$ *such that* $u(\cdot, t)$ *is discontinuous for any* $t < T^*$*, while it is continuous for any* $t > T^*$.

Example. Let us define $u_l = -2$, $u_r = 2$, $f(u) = (u^2 - 1)^2$, and $g(u) = 4 - u^2$. Then the assumptions of Theorem 2.5 are satisfied with

$$\bar{u}_1 = -\frac{1}{\sqrt{3}}, \qquad \bar{u}_2 = \frac{1}{\sqrt{3}}, \qquad \hat{v}_1 = -1, \qquad \hat{v}_2 = 1.$$

3. Preliminaries. It is well known (see, for instance, [10]) that under our assumptions, problem (1.1)–(1.2) possesses a unique solution in the class of the so-called *entropy solutions*, taking values in $[u_m, u_M]$. For our purposes, is enough to recall the following characterization of piecewise-smooth entropy solutions (see [8]). It is convenient to use the notation

$$\sigma(u,v) = \begin{cases} \dfrac{f(u) - f(v)}{u - v} & \text{if } u \neq v, \\[2ex] f'(u) & \text{if } u = v \end{cases} \tag{3.1}$$

for $u, v \in [u_m, u_M]$.

Theorem 3.1. *Suppose that* $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ *is of class* C^1 *in the complement of a finite number of smooth curves* $x = \gamma_1(t), \ldots, x = \gamma_k(t)$ *which intersect themselves at most at a finite number of points, and suppose that the one-sided limits of* u *exist along* $\gamma_1, \ldots, \gamma_k$. *Then* u *is an entropy solution of problem* (1.1)–(1.2) *if and only if the following hold.*

(i) *Equalities* (1.1) *and* (1.2) *are satisfied in the classical sense in the complement of the curves* γ_i.

(ii) *Along each curve* γ_i*, the function* u *satisfies the* Rankine–Hugoniot jump condition

$$\gamma_i'(t)[u_l(t) - u_r(t)] = [f(u_l(t)) - f(u_r(t))] \quad \forall\, t > 0 \tag{3.2}$$

and the Oleinik admissibility condition

$$\begin{gathered} \sigma(u_l(t), v) \geq \sigma(u_l(t), u_r(t)) \geq \sigma(u_r(t), v) \\ \forall\, t > 0, \quad \forall\, v \in [\min\{u_l(t), u_r(t)\}, \max\{u_l(t), u_r(t)\}]. \end{gathered} \tag{3.3}$$

Here we have set

$$u_l(t) = \lim_{x \to \gamma_i(t)^-} u(x,t),$$

$$u_r(t) = \lim_{x \to \gamma_i(t)^+} u(x,t).$$

Remark 3.2. The two inequalities in (3.3) are equivalent since $\sigma(u_l(t), u_r(t))$ is a convex combination of $\sigma(u_l(t), v)$ and $\sigma(u_r(t), v)$.

To find the entropy solution of problem (1.1)–(1.2), we will construct suitable piecewise-smooth functions and then check that they satisfy conditions (i) and (ii) of Theorem 3.1. To this end, it is useful to introduce some auxiliary functions.

Definition 3.3. *For any $v \in [u_m, u_M]$, let $W(v,\cdot)$ denote the solution of*

$$\begin{cases} \partial_t W(v,t) = g(W(v,t)), \quad t \in \mathbb{R}, \\ W(v,0) = v. \end{cases} \tag{3.4}$$

For $v \in [u_m, u_M]$ and $t \geq 0$, let us also set

$$F_+(v,t) = \int_0^t f'(W(v,s))\,ds, \tag{3.5}$$

$$F(v,t) = \int_0^t f'(W(v,-s))\,ds. \tag{3.6}$$

The following result motivates the definition of W, F, and F_+ and collects some useful properties of such functions.

Lemma 3.4.

(i) *Given $u_o \in [u_m, u_M]$ and $x_o \in \mathbb{R}$, the functions*

$$\begin{cases} x(t) = x_o + F_+(u_o, t), \\ u(t) = W(u_o, t), \end{cases} \qquad t \geq 0, \tag{3.7}$$

are the solution of the characteristic system *associated with equation* (1.1)*,*

$$\begin{cases} x'(t) = f'(u(t)), \\ u'(t) = g(u(t)) \end{cases} \tag{3.8}$$

with initial data $x(0) = x_o$, $u(0) = u_o$. Similarly, given $\bar{u} \in [u_m, u_M]$ and $(\bar{x}, \bar{t}) \in \mathbb{R} \times \mathbb{R}_+$, the functions

$$\begin{cases} x(t) = \bar{x} - F(\bar{u}, \bar{t} - t), \\ u(t) = W(\bar{u}, t - \bar{t}), \end{cases} \qquad t \in [0, \bar{t}], \tag{3.9}$$

are the solution of system (3.8) *with terminal conditions $x(\bar{t}) = \bar{x}$, $u(\bar{t}) = \bar{u}$.*

(ii) *If $v = u_l$ or $v = u_r$, then $W(v,t) \equiv v$ and $F_+(v,t) = F(v,t) = f'(v)t$.*

(iii) *If $v \neq u_l, u_r$, then $W(v,\cdot)$ is strictly increasing and F and F_+ satisfy*

$$F(v,t) = F_+(W(v,-t),t). \tag{3.10}$$

(iv) *F is of class C^1 and satisfies*

$$\partial_t F(v,t) = f'(W(v,-t)), \tag{3.11}$$

$$\partial_v F(v,t) = \frac{f'(v) - f'(W(v,-t))}{g(v)} \quad (v \neq u_l, u_r). \tag{3.12}$$

(v) *Let A be an open subset of $\mathbb{R} \times \mathbb{R}_+$, and let $\tilde{u} \in C^1(A)$ satisfy*

$$F(\tilde{u}(x,t),t) = x \quad \textit{for any } (x,t) \in A. \tag{3.13}$$

Then $\tilde{u}$ solves equation (1.1) *in A.*

Proof. Properties (i) to (iii) are easy consequences of the definition and of assumption (H3) as well as equality (3.11). Let us check (3.12). For $v \neq u_l, u_r$, we have

$$\begin{aligned}\partial_v W(v,t) &= \exp\left(\int_0^t g'(W(v,s))ds\right)\\ &= \exp\left(\int_v^{W(v,t)} \frac{g'(w)}{g(w)}dw\right) = \frac{g(W(v,t))}{g(v)},\end{aligned}$$

$$\begin{aligned}\partial_v F(v,t) &= \int_0^t f''(W(v,-s))\frac{g(W(v,-s))}{g(v)}\,ds\\ &= -\frac{1}{g(v)}\int_0^t \frac{d}{ds}f'(W(v,-s))ds\\ &= \frac{f'(v) - f'(W(v,-t))}{g(v)}.\end{aligned}$$

To verify that property (v) holds, let us first observe that by equalities (3.11) and (3.12),

$$\partial_t F(v,t) + g(v)\partial_v F(v,t) = f'(v). \tag{3.14}$$

Differentiating (3.13), we obtain

$$\partial_v F\,\partial_x \tilde{u} = 1, \qquad \partial_v F\,\partial_t \tilde{u} + \partial_t F = 0.$$

Thus

$$\begin{aligned}0 &= \partial_x \tilde{u}\,(\partial_v F\,\partial_t \tilde{u} + \partial_t F) = \partial_t \tilde{u} + \partial_x \tilde{u}\,\partial_t F\\ &= \partial_t \tilde{u} + f'(\tilde{u})\partial_x \tilde{u} - g(\tilde{u}). \qquad \square\end{aligned}$$

Remark 3.5. Some comments about equality (3.13) are in order. Since we are considering Riemann initial data with a discontinuity which is not entropy admissible, we expect the solution to exhibit a fan of characteristics starting from the origin. Let us therefore consider a solution $\tilde{u}$ of (1.1) and assume that the backward characteristic from a given point (x,t) ends up at the origin. By (3.9), the value $\tilde{u}(x,t)$ satisfies equality (3.13). Thus such an equality characterizes the solutions defined by a fan of characteristics centered at the origin.

This fact is well known for the case where $g \equiv 0$. In fact, defining the functions W and F as above, we obtain $W(v,t) = v$ and $F(v,t) = f'(v)t$. Therefore, (3.13) becomes

$$f'(\tilde{u}(x,t)) = \frac{x}{t},$$

which is the equality that characterizes the rarefaction waves centered at the origin. This equality is usually derived by exploiting the self-similarity of the Riemann problem under the transformation $(x,t) \to (\lambda x, \lambda t)$ for $\lambda > 0$ rather than the method of characteristics. Such rescaling arguments, however, cannot be applied to the general inhomogeneous case.

In the following sections, we will use equality (3.13) to define $\tilde{u}$. This is equivalent to finding the inverse of F with respect to the variable v. The problem is that $\partial_v F$ is not of constant sign by (3.12) and (H2). Therefore, equality (3.13) in general defines a multivalued function, and every time, we have to select the branch of $\tilde{u}$ which fits to our purposes. This is a major complication due to the nonconvexity of f when we want to study equation (1.1) with the method of characteristics. Let us also observe that in the homogeneous case, the values of v for which $\partial_v F(v,t) = 0$ are the zeros of f'' and do not depend on t. In our case, because of the presence of the source term, these values vary with time (and so do the restrictions to be imposed on the range of $\tilde{u}$ in order to obtain a single-valued function from (3.13)).

4. The case where $u_l < u_r$.

Throughout this section, we assume that properties (H1)–(H4) hold and that $u_l < u_r$.

LEMMA 4.1. *There exists a unique $\hat{v} \in \,]\bar{u}, u_r[$ such that*

$$f'(\hat{v}) = \sigma(u_l, \hat{v}). \tag{4.1}$$

Moreover, the function $v \to \sigma(u_l, v)$ is decreasing for $v \in [u_l, \hat{v}]$.

Proof. By assumptions (H2) and (H4), we have

$$f'(\bar{u}) < \sigma(u_l, \bar{u}), \qquad f'(u_r) > \sigma(u_l, u_r).$$

Therefore, $f'(v) = \sigma(u_l, v)$ for some $v \in \,]\bar{u}, u_r[$. Denote by $\hat{v}$ the smallest of the values satisfying this equality. Then for any $v > \hat{v}$,

$$\begin{aligned}\sigma(u_l, v) &= (v - u_l)^{-1}\left(\int_{u_l}^{\hat{v}} f'(w)dw + \int_{\hat{v}}^{v} f'(w)dw\right)\\ &= (v - u_l)^{-1}\left((\hat{v} - u_l)f'(\hat{v}) + \int_{\hat{v}}^{v} f'(w)dw\right) < f'(v)\end{aligned}$$

because, by (H2), f' is increasing in $[\bar{u}, u_r]$. It follows that there is a unique value that satisfies (4.1).

Differentiating σ, we obtain

$$\partial_v \sigma(u_l, v) = \frac{f'(v) - \sigma(u_l, v)}{v - u_l}.$$

Since we have $f'(v) < \sigma(u_l, v)$ for any $v \in \,]u_l, \hat{v}[\,$, we deduce that $\sigma(u_l, \cdot)$ is decreasing in $[u_l, \hat{v}]$. ☐

DEFINITION 4.2. *Let $\hat{v}$ be the value given by Lemma* 4.1.

(i) *For any (x, t) with $t > 0$ and $x \in [F_+(\hat{v}, t), f'(u_r)t]$, let $\tilde{u}(x, t) \in [W(\hat{v}, t), u_r]$ be defined by*

$$F(\tilde{u}(x, t), t) = x.$$

(ii) *Let $\phi : [0, \infty[\,\to [\hat{v}, u_r[$ be defined by*

$$\int_{\hat{v}}^{\phi(\xi)} \frac{f'(v) - f'(\hat{v})}{g(v)}\, dv = \xi \quad \textit{for } \xi \geq 0.$$

LEMMA 4.3.

(i) *$\tilde{u}$ is well defined, of class C^1, and solves equation* (1.1) *in the interior of its domain of definition.*

(ii) *ϕ is well defined, belongs to $C([0, \infty[\,) \cap C^2(\,]0, \infty[\,)$, and satisfies*

$$\begin{cases} \phi' = \dfrac{g(\phi)}{f'(\phi) - f'(\hat{v})} & \textit{in }]0, \infty[, \\[2ex] \phi(0) = \hat{v}. \end{cases} \tag{4.2}$$

(iii) *For any $t > 0$, we have*

$$\tilde{u}(F_+(\hat{v}, t), t) = \phi(F_+(\hat{v}, t) - f'(\hat{v})t) = W(\hat{v}, t),$$

$$\tilde{u}(f'(u_r)t, t) = u_r.$$

Proof. (i) Let us fix $t > 0$ and $v \in [W(\hat{v}, t), u_r]$. Then by assumption (H3) and the definition of W,

$$u_r \geq v > W(v, -t) \geq \hat{v}.$$

Since $\hat{v} > \bar{u}$ and f' is increasing in $[\bar{u}, u_r]$ by assumption (H2), we have

$$f'(v) > f'(W(v, -t)).$$

Therefore, by (3.12),

$$\partial_v F(v, t) > 0 \quad \forall\, v \in [W(\hat{v}, t), u_r[.$$

Since, by Lemma 3.4(ii)–(iii), $F(W(\hat{v}, t), t) = F_+(\hat{v}, t)$ and $F(u_r, t) = f'(u_r)t$, we obtain that $\tilde{u}$ is well defined, of class C^1, and satisfies

$$\tilde{u}(F_+(\hat{v}, t), t) = W(\hat{v}, t), \tag{4.3}$$

$$\tilde{u}(f'(u_r)t, t) = u_r. \tag{4.4}$$

Furthermore, by Lemma 3.4(v), $\tilde{u}$ satisfies equation (1.1).

(ii) It follows from the definition and from assumptions (H2) and (H3).

(iii) By Definition 3.3,

$$F_+(\hat{v},t) - f'(\hat{v})t = \int_0^t [f'(W(\hat{v},s)) - f'(\hat{v})]\,ds$$
$$= \int_{\hat{v}}^{W(\hat{v},t)} \frac{f'(w) - f'(\hat{v})}{g(w)}\,dw.$$

Therefore,

$$W(\hat{v},t) = \phi(F_+(\hat{v},t) - f'(\hat{v})t).$$

We conclude (4.4) from (4.3). □

Theorem 4.4. *Let assumptions* (H1)–(H4) *be satisfied and let* $u_l < u_r$. *Then the solution of problem* (1.1)–(1.2) *is*

$$(4.5)\qquad u(x,t) = \begin{cases} u_l, & x < f'(\hat{v})t, \\ \phi(x - f'(\hat{v})t), & f'(\hat{v})t < x < F_+(\hat{v},t), \\ \tilde{u}(x,t), & F_+(\hat{v},t) \le x \le f'(u_r)t, \\ u_r, & x > f'(u_r)t, \end{cases}$$

where $\hat{v}$, $\tilde{u}$, *and* ϕ *are defined as in Lemma* 4.1 *and Definition* 4.2.

Proof. It is easily checked that the function defined by (4.5) satisfies the hypotheses of Theorem 3.1. In fact, by Lemma 4.3(i)–(ii) and by assumption (H3), u is a classical solution of equation (1.1) in each of the four regions which appear in (4.5). By Lemma 4.3(iii), u is continuous along the curves $x = F_+(\hat{v},t)$ and $x = f'(u_r)t$. Finally, Lemma 4.1 implies that conditions (3.2) and (3.3) are satisfied along the line of discontinuity $x = f'(\hat{v})t$ (see also Remark 3.2). □

Proof of Theorem 2.1. Let us fix $t > 0$ and $x > F_+(\hat{v},t)$. By Lemma 4.3(iii),

$$u_r > \phi(x - f'(\hat{v})t) > \phi(F_+(\hat{v},t) - f'(\hat{v})t) = W(\hat{v},t).$$

On the other hand, by Theorem 4.4 and Definition 4.2,

$$u_r \ge u(x,t) \ge W(\hat{v},t).$$

It follows that

$$|u(x,t) - \phi(x - f'(\hat{v})t)| \le u_r - W(\hat{v},t) \quad \forall\, t > 0, \quad x > F_+(\hat{v},t).$$

Since, by assumption (H3), $W(\hat{v},t) \to u_r$ as $t \to \infty$, the conclusion follows. □

5. The case where $u_l > u_r$. Throughout this section, we assume that properties (H1)–(H4) hold and that $u_l > u_r$. The following result is analogous to Lemma 4.1.

Lemma 5.1. *There exists a unique* $\hat{w} \in \,]u_r, \bar{u}[$ *such that*

$$(5.1)\qquad f'(\hat{w}) = \sigma(\hat{w}, u_l).$$

Moreover, the function $v \to \sigma(v, u_l)$ *is increasing for* $v \in [\hat{w}, u_l]$.

As observed in Remark 3.5, it is useful to find the values for which $\partial_v F(v,t) = 0$. This is done in the next lemma.

LEMMA 5.2. *For fixed $t > 0$, there exists a unique value $v^* = v^*(t) \in \,]\bar{u}, u_l[$ such that*

$$f'(v^*) = f'(W(v^*, -t)). \tag{5.2}$$

The function $t \to v^*(t)$ is strictly increasing for $t > 0$ and satisfies

$$\lim_{t\to 0} v^*(t) = \bar{u}, \tag{5.3}$$

$$\lim_{t\to\infty} v^*(t) = \begin{cases} u_l & \text{if } f'(u_r) \geq f'(u_l), \\ \bar{v} & \text{if } f'(u_r) < f'(u_l), \end{cases} \tag{5.4}$$

where $\bar{v}$ is the unique value in $]u_r, u_l[$ satisfying $f'(\bar{v}) = f'(u_r)$. Moreover,

$$\text{sgn}\partial_v F(v,t) = \text{sgn}[v - v^*(t)] \tag{5.5}$$

for any $t > 0$ and $v \in \,]u_r, u_l[$.

Proof. Let us fix $t > 0$. Since g is positive, we have by (3.4) that $v > W(v, -t)$ for any $v \in \,]u_r, u_l[\,$. Using property (H2), we obtain that

$$f'(v) < f'(W(v,-t)) \quad \text{for any } v \in \,]u_r, \bar{u}], \tag{5.6}$$

$$f'(v) > f'(W(v,-t)) \quad \text{for any } v \in [W(\bar{u}, t), u_l[\,. \tag{5.7}$$

Therefore, there exists at least one value v^* for which equality (5.2) holds. Moreover, any such v^* satisfies

$$v^* \in \,]\bar{u}, W(\bar{u}, t)[, \quad W(v^*, -t) \in \,]W(\bar{u}, -t), \bar{u}[\,. \tag{5.8}$$

Since $f'(v)$ is decreasing for $v \in [W(\bar{u}, -t), \bar{u}]$ and increasing for $v \in [\bar{u}, W(\bar{u}, t)]$, the value v^* satisfying (5.2) is unique. From inequalities (5.6) and (5.7), it follows that

$$\text{sgn}[f'(v) - f'(W(v,-t))] = \text{sgn}[v - v^*]. \tag{5.9}$$

This implies (5.5) by Lemma 3.4(iv) and assumption (H3).

Let us now take $t_2 > t_1 > 0$. We have

$$W(v^*(t_1), -t_2) < W(v^*(t_1), -t_1) < \bar{u}.$$

Since $f'(v)$ is decreasing for $v \leq \bar{u}$, we obtain

$$f'(W(v^*(t_1), -t_2)) > f'(W(v^*(t_1), -t_1)) = f'(v^*(t_1)),$$

which implies by (5.9) that $v^*(t_2) > v^*(t_1)$. Hence v^* is strictly increasing. In a similar way, it is checked that $W(v^*(t), -t)$ is a decreasing function of t.

Property (5.3) follows from (5.8) since $W(\bar{u}, 0) = \bar{u}$. To prove (5.4), let us observe that by the definition of W,

$$\int_{W(v^*(t),-t)}^{v^*(t)} \frac{dv}{g(v)} = t. \tag{5.10}$$

Let us denote by v_∞ and w_∞ the limits of $v^*(t)$ and $W(v^*(t),-t)$ as $t \to \infty$. By relations (5.2), (5.8), and (5.10), these values satisfy

$$f'(w_\infty) = f'(v_\infty), \tag{5.11}$$

$$u_r \leq w_\infty < v_\infty \leq u_l, \tag{5.12}$$

$$\int_{w_\infty}^{v_\infty} \frac{dv}{g(v)} = \infty. \tag{5.13}$$

From (5.12) and (5.13), we deduce that either $w_\infty = u_r$ or $v_\infty = u_l$. Then equality (5.11) and assumption (H2) yield (5.4). ∎

Definition 5.3. *For any $t > 0$ and $x \in [F(v^*(t),t), f'(u_r)t]$, let $\tilde{u}(x,t) \in [u_r, v^*(t)]$ be defined by*

$$F(\tilde{u}(x,t),t) = x. \tag{5.14}$$

Remark 5.4. In this paper, we sometimes denote different functions with the same symbol because they play the same role in the problem that is being considered (in this case, the function $\tilde{u}$; see Definition 4.2). However, there is no ambiguity since each of these functions appears only in the section where it has been introduced.

Lemma 5.5. *$\tilde{u}$ is well defined. It is of class C^1 and satisfies equation* (1.1) *in the interior of its domain of definition. Furthermore,*

$$\tilde{u}(f'(u_r)t,t) = u_r \quad \forall\, t > 0. \tag{5.15}$$

Proof. By Lemma 3.4(ii), we have $F(u_r,t) = f'(u_r)t$. Equality (5.5) ensures that $F(\cdot,t)$ is strictly decreasing in $[u_r, v^*(t)]$. Therefore, $\tilde{u}$ is well defined and satisfies (5.15). From Lemma 3.4(v), it follows that $\tilde{u}$ solves equation (1.1). ∎

Lemma 5.6. *There exists $\gamma \in C^1([0,\infty[)$ with the following properties.*

$$\gamma(0) = 0, \qquad \gamma'(0) = f'(\hat{w}), \tag{5.16}$$

$$\gamma(t) \in \,]F(v^*(t),t), F(\hat{w},t)[\quad \forall t > 0, \tag{5.17}$$

$$\tilde{u}(\gamma(t),t) \in \,]\hat{w}, v^*(t)[\quad \forall t > 0, \tag{5.18}$$

$$\gamma'(t) = \sigma(\tilde{u}(\gamma(t),t), u_l) \quad \forall t > 0, \tag{5.19}$$

$$\tilde{u}(\gamma(\cdot),\cdot) \text{ is increasing in }]0,\infty[\,, \tag{5.20}$$

$$\lim_{t\to 0} \tilde{u}(\gamma(t),t) = \hat{w}, \tag{5.21}$$

$$\lim_{t\to\infty} \tilde{u}(\gamma(t),t) = \begin{cases} u_l & \text{if } f'(u_r) \geq f'(u_l), \\ \bar{w} & \text{if } f'(u_r) < f'(u_l), \end{cases} \tag{5.22}$$

where $\bar{w} \in]\hat{w}, u_l[$ *is the unique value satisfying* $f'(u_r) = \sigma(u_l, \bar{w})$.

Remark 5.7. In analogy with the homogeneous case (see Theorem 2.4), we want to show that the states u_l and u_r can be connected by a shock wave followed by a rarefaction wave. The shock curve is given by $x = \gamma(t)$, where γ is the function whose existence we are going to prove, while the rarefaction wave is given by the function $\tilde{u}$. According to condition (3.2), the speed of γ is given by (5.19). In order to stay within the domain of definition of $\tilde{u}$, we must have $\gamma(t) \geq F(v^*(t), t)$. On the other hand, we cannot have $\gamma(t) > F(\hat{w}, t)$; otherwise, $\tilde{u}(\gamma(t), t))$ would assume a value smaller than $\hat{w}$, which could not be connected to u_l by an admissible shock. This motivates the restrictions in (5.17). We also notice that by property (5.16), for small times, $\gamma(t)$ is close to $f'(\hat{w})t$, which is the position of the shock in the homogeneous case.

Proof of Lemma 5.6. To simplify notation, we set

$$\chi_1(t) = F(v^*(t), t),$$

$$\chi_2(t) = F(\hat{w}, t),$$

$$h(x, t) = \sigma(\tilde{u}(x, t), u_l).$$

We have to show that there exists $\gamma \in C^1([0, \infty[)$ such that $\chi_1(t) < \gamma(t) < \chi_2(t)$ for any $t > 0$ and

$$\begin{cases} \gamma'(t) = h(\gamma(t), t) \quad \forall\, t > 0, \\ \gamma(0) = 0. \end{cases} \tag{5.23}$$

Let us observe that by equalities (3.11), (3.12), (5.1), (5.2), and (5.14),

$$\chi_1'(t) = f'(v^*(t)), \qquad \chi_2'(t) = f'(W(\hat{w}, -t)), \tag{5.24}$$

$$h(\chi_1(t), t) = \sigma(v^*(t), u_l), \qquad h(\chi_2(t), t) = f'(\hat{w}). \tag{5.25}$$

We recall that by Lemmas 5.1 and 5.2,

$$u_r < \hat{w} < \bar{u} < v^*(t) < u_l \quad \forall\, t > 0.$$

From assumption (H2) and inequalities (5.24) and (5.25), we deduce

$$\chi_1'(t) < h(\chi_1(t), t), \qquad \chi_2'(t) > h(\chi_2(t), t) \quad \forall\, t > 0. \tag{5.26}$$

We now consider equation (5.23). Since $\chi_1(0) = \chi_2(0) = 0$ and h is not defined at $(0, 0)$, we will prove the existence of a solution by an approximation procedure. We first observe that by (5.24), (5.3), and (H2), we have

$$\chi_1'(t) < f'(\hat{w})$$

for t small. On the other hand, since $W(\hat{w}, \cdot)$ is increasing, we obtain by the same relations that

$$\chi_2'(t) > f'(\hat{w})$$

for any $t > 0$. It follows that for $n \in \mathbb{N}$ large enough,

$$\chi_1\left(\frac{1}{n}\right) < \frac{f'(\hat{w})}{n} < \chi_2\left(\frac{1}{n}\right).$$

Define γ_n to be the solution of

$$\begin{cases} \gamma_n'(t) = h(\gamma_n(t), t) & \text{for } t \geq \dfrac{1}{n} \\ \gamma_n\left(\dfrac{1}{n}\right) = \dfrac{f'(\hat{w})}{n}. \end{cases}$$

The inequalities in (5.26) show that γ_n is well defined for $t \in [1/n, \infty[$ and satisfies $\chi_1 < \gamma_n < \chi_2$. Since, by Lemma 5.1 and Definition 5.3,

$$h(x,t) > \sigma(\hat{w}, u_l) = f'(\hat{w}) \quad \forall\, x \in [\chi_1(t), \chi_2(t)[\,,$$

we also have

$$\gamma_n(t) > f'(\hat{w})t \quad \forall\, t > \frac{1}{n}.$$

It follows that

$$\max\{\chi_1(t), f'(\hat{w})t\} < \gamma_n(t) < \gamma_{n+1}(t) < \chi_2(t)$$

for any n, $t > 1/n$. Thus if we define

$$\gamma(t) = \begin{cases} \lim\limits_{n\to\infty} \gamma_n(t) & \text{if } t > 0, \\ 0 & \text{if } t = 0, \end{cases}$$

it is easily seen that γ satisfies (5.17) and (5.19). (The strict inequality in (5.17) follows from (5.26).) Property (5.18) is equivalent to (5.17) by Definition 5.3. We have also obtained that

$$\gamma(t) > f'(\hat{w})t \quad \text{for any } t > 0. \tag{5.27}$$

Let us now fix $v' \in]\hat{w}, \bar{u}]$. Since, by Lemma 3.4(iv),

$$\lim_{t\to 0} \partial_t F(v', t) = f'(v') < f'(\hat{w}),$$

there exists $\varepsilon = \varepsilon(v')$ such that

$$F(v', t) < f'(\hat{w})t \quad \forall\, t < \varepsilon.$$

By the definition of $\tilde{u}$ and of χ_2, this implies

$$\tilde{u}(x,t) \in [\hat{w}, v'[\quad \forall\, t < \varepsilon, \quad x \in [f'(\hat{w})t, \chi_2(t)].$$

Since v' can be chosen arbitrarily close to $\hat{w}$, we obtain

$$\lim_{\substack{(x,t)\to(0,0) \\ f'(\hat{w})t \leq x \leq \chi_2(t)}} \tilde{u}(x,t) = \hat{w},$$

$$\lim_{\substack{(x,t)\to(0,0)\\ f'(\hat{w})t\le x\le \chi_2(t)}} h(x,t) = \sigma(\hat{w},u_l) = f'(\hat{w}).$$

From (5.17) and (5.27), we deduce that

$$\lim_{t\to 0} \tilde{u}(\gamma(t),t) = \hat{w}, \qquad \lim_{t\to 0} \gamma'(t) = f'(\hat{w}).$$

Thus we have proved properties (5.16) and (5.21).

For simplicity, let us set

$$\alpha(t) = \tilde{u}(\gamma(t),t).$$

To show that (5.20) holds, let us first observe that by (5.14) and (5.5),

$$\text{sgn}\,[\alpha(t) - v] = \text{sgn}\,[F(v,t) - \gamma(t)] \tag{5.28}$$

for any $t > 0$ and $v \in [u_r, v^*(t)]$. Suppose that $\alpha(\cdot)$ is not increasing. Then there exist $t_2 > t_1 > 0$ such that $\alpha(t_2) < \alpha(t_1)$. If we fix $v_0 \in]\alpha(t_2), \alpha(t_1)[$, we deduce by (5.28) that there exists $t_3 \in]t_1, t_2[$ such that

$$\alpha(t_3) = v_0, \qquad \gamma'(t_3) \ge \partial_t F(v_0, t_3). \tag{5.29}$$

Since $\alpha(t_1) > v_0 > \hat{w} = \alpha(0)$, we also obtain that there exists $t_4 \in]0, t_1[$ such that

$$\alpha(t_4) = v_0, \qquad \gamma'(t_4) \le \partial_t F(v_0, t_4). \tag{5.30}$$

From relations (3.11), (5.29), and (5.30), we deduce

$$f'(W(v_0), -t_4) \ge f'(W(v_0), -t_3). \tag{5.31}$$

Since $t_4 < t_3$, we have $W(v_0, -t_4) > W(v_0, -t_3)$. Moreover, by inclusions (5.18) and (5.8), $W(v_0, -t_4) < W(v^*(t_4), -t_4) < \bar{u}$. However, f' is decreasing in $[u_r, \bar{u}]$, and thus we find a contradiction to (5.31). This proves that $\alpha(\cdot)$ is increasing.

It remains to prove (5.22). Let us denote by w_∞ the limit of $\alpha(t)$ as $t \to \infty$. By (5.19), we have

$$\lim_{t\to\infty} \gamma'(t) = \sigma(w_\infty, u_l). \tag{5.32}$$

Suppose that $f'(u_r) \ge f'(u_l)$ and $w_\infty < u_l$. Then by (3.11) and Lemma 5.1,

$$\lim_{t\to\infty} \partial_t F(v,t) = \lim_{t\to\infty} f'(W(v,-t)) = f'(u_r) \ge f'(u_l) > \sigma(w_\infty, u_l)$$

for any $v < u_l$. By (5.32), this implies that

$$\gamma(t) < F(v,t) \quad \forall\, t \gg 0.$$

By (5.28), we obtain

$$\alpha(t) > v \quad \forall\, t \gg 0.$$

If we choose $v > w_\infty$, we find a contradiction to the definition of w_∞. Thus relation (5.22) is proved in the case $f'(u_r) \ge f'(u_l)$.

Let us now suppose that $f'(u_r) < f'(u_l)$. In this case, Lemma 5.2 asserts that $\lim_{t\to\infty} v^*(t) < u_l$. Then from (5.18), we obtain that $w_\infty < u_l$. By the same argument used in the case where $f'(u_r) \geq f'(u_l)$, we find a contradiction unless

$$f'(u_r) = \sigma(w_\infty, u_l).$$

By Lemma 5.1, there exists only one value $w_\infty \in]\hat{w}, u_l[$ satisfying the above equality. The proof is complete. ◻

THEOREM 5.8. *Suppose that assumptions* (H1)–(H4) *are satisfied and that* $u_l > u_r$. *Let* $\tilde{u}$ *be defined as in Definition* 5.3. *Then the function* γ *given by Lemma* 5.6 *is unique and the solution of problem* (1.1)–(1.2) *is*

$$u(x,t) = \begin{cases} u_l, & x < \gamma(t), \\ \tilde{u}(x,t), & \gamma(t) < x \leq f'(u_r)t, \\ u_r, & x > f'(u_r)t. \end{cases} \tag{5.33}$$

Proof. By Lemma 5.5 and assumption (H3), u is a classical solution of equation (1.1) in each of the three regions defined above. It is continuous on the line $x = f'(u_r)t$ by virtue of (5.15). Furthermore, from (5.19), (5.18), and Lemma 5.1, it follows that along the curve $x = \gamma(t)$, conditions (3.2) and (3.3) are satisfied. We conclude by Theorem 3.1 that u is the solution of our problem. The uniqueness of γ follows from the uniqueness of u. ◻

Proof of Theorem 2.3.

Step 1. We have

$$\lim_{t\to\infty} \gamma(t) - f'(u_r)t = \begin{cases} -L & \text{in cases (i) and (ii)}, \\ -\infty & \text{in case (iii)}. \end{cases} \tag{5.34}$$

Let us first prove this property for cases (i) and (ii). From the proof of Lemma 5.2, it follows that

$$\lim_{t\to\infty} W(v^*(t), -t) = u_r. \tag{5.35}$$

For simplicity, we set $\alpha(t) := \tilde{u}(\gamma(t), t)$, $\omega(t) := W(\alpha(t), -t)$. By (5.18),

$$u_r < \omega(t) < W(v^*(t), -t).$$

Therefore,

$$\lim_{t\to\infty} \omega(t) = u_r. \tag{5.36}$$

On the other hand, by (5.22),

$$\lim_{t\to\infty} \alpha(t) = \begin{cases} u_l & \text{if } f'(u_r) = f'(u_l), \\ \bar{w} & \text{if } f'(u_r) < f'(u_l). \end{cases} \tag{5.37}$$

From (5.14), (3.4), and (3.6), we deduce that

$$\gamma(t) - f'(u_r)t = F(\alpha(t), t) - f'(u_r)t$$

$$= \int_0^t [f'(W(\alpha(t), -s)) - f'(u_r)]ds \tag{5.38}$$

$$= \int_{\omega(t)}^{\alpha(t)} \frac{f'(v) - f'(u_r)}{g(v)}\, dv.$$

By letting $t \to \infty$ and taking into account relations (5.36) and (5.37), we prove the first part of (5.34). Let us also observe that by (5.20), equality (5.38) implies that

$$\gamma(t) > f'(u_r)t - L \quad \forall\, t > 0. \tag{5.39}$$

To prove (5.34) in case (iii) it suffices to observe that by (5.19) and (5.22),

$$\lim_{t\to\infty} \gamma'(t) = f'(u_l) < f'(u_r).$$

Step 2. Let us fix $\xi \in\,]-L, 0]$ (in cases (i) or (ii)) or $\xi \in\,]-\infty, 0]$ (in case (iii)) and define

$$\lambda(t) = \lambda_\xi(t) = \tilde{u}(f'(u_r)t + \xi, t).$$

Then $\lambda(\cdot)$ is decreasing and

$$\lim_{t\to\infty} \lambda(t) = \psi(\xi).$$

Let us first show that λ is well defined for t large enough. From Step 1 and (5.17), we deduce that

$$F(v^*(t), t) < \gamma(t) < f'(u_r)t + \xi \le f'(u_r)t \quad \forall\, t \gg 0,$$

and therefore $\tilde{u}$ is defined at $(f'(u_r)t + \xi, t)$. Next, we observe that by the definition of $\tilde{u}$,

$$F(\lambda(t), t) - f'(u_r)t \equiv \xi. \tag{5.40}$$

Differentiating this equality with respect to t, we obtain by (3.11) that

$$\partial_v F(\lambda(t), t)\lambda'(t) + f'(W(\lambda(t), -t)) - f'(u_r) \equiv 0. \tag{5.41}$$

Since $\lambda(t) \in [u_r, v^*(t)[$, we obtain from (3.12), (5.5), (5.8), and (H2) that

$$\partial_v F(\lambda(t), t) < 0, \qquad f'(W(\lambda(t), -t)) \le f'(u_r).$$

Thus equality (5.41) implies that $\lambda(\cdot)$ is decreasing. Let us denote by λ_∞ its limit as $t \to \infty$. From (5.40), (3.4), and (3.6), we deduce that

$$\begin{aligned} \xi &= \int_0^t [f'(W(\lambda(t), -s)) - f'(u_r)]\, ds \\ &= \int_{W(\lambda(t), -t)}^{\lambda(t)} \frac{f'(v) - f'(u_r)}{g(v)}\, dv \\ &\overset{t\to\infty}{\longrightarrow} \int_{u_r}^{\lambda_\infty} \frac{f'(v) - f'(u_r)}{g(v)}\, dv. \end{aligned}$$

Thus $\lambda_\infty = \psi(\xi)$.

Step 3. $\sup\{|\tilde{u}(x,t) - \psi(x - f'(u_r)t)| \;:\; x \in [\gamma(t), f'(u_r)t]\} \to 0$ as $t \to \infty$. It is convenient to set

$$\bar{z} = \lim_{t\to\infty} \alpha(t) = \begin{cases} u_l & \text{in cases (i) and (iii)}, \\ \bar{w} & \text{in case (ii)}. \end{cases}$$

For fixed $\varepsilon \in \,]0, \bar{z} - u_r[$, let us define

$$\xi_\varepsilon = \int_{u_r}^{\bar{z}-\varepsilon} \frac{f'(v) - f'(u_r)}{g(v)}\, dv.$$

By Step 1, we have

$$\gamma(t) < f'(u_r)t - \xi_\varepsilon \quad \forall\, t \gg 0.$$

By Step 2, we can apply Dini's theorem to obtain

$$\lim_{t\to\infty} \tilde{u}(\xi + f'(u_r)t, t) = \psi(\xi), \tag{5.42}$$

uniformly for $\xi \in [\xi_\varepsilon, 0]$. Let us now fix T such that

$$\tilde{u}(\xi_\varepsilon + f'(u_r)t, t) - \psi(\xi_\varepsilon) < \varepsilon \quad \forall\, t \geq T.$$

Then, since $\tilde{u}(\cdot, t)$ and $\psi(\cdot)$ are both decreasing and since inequality (5.39) holds, we obtain for $\xi \in [\gamma(t) - f'(u_r)t, \xi_\varepsilon]$ and $t \geq T$ that

$$\psi(\xi) \in [\bar{z} - \varepsilon, \bar{z}], \qquad \tilde{u}(\xi + f'(u_r)t, t) \in [\bar{z} - 2\varepsilon, \alpha(t)]. \tag{5.43}$$

Step 3 then follows from relations (5.42) and (5.43). From Steps 1 and 3 and from Theorem 5.8, it is easy to deduce Theorem 2.3. □

6. Extinction of shocks in finite time. In the previous sections, we have considered the case when f has a single inflection point between u_l and u_r and g has constant sign. It is possible to treat more complicated situations with a similar procedure and obtain an asymptotic profile for the solution consisting of a suitable number of shock waves and traveling waves. An interesting new property arising when f has more than one inflection point is that the discontinuities present in the solution may vanish in finite time. In this section, we give an example of such a behavior.

We assume that $u_l < u_r$ and that f and g satisfy hypotheses (H1), (H3), (H5), (H6), and (H7).

LEMMA 6.1. *For any $v \in [\hat{v}_1, \bar{u}_2]$, there exists a unique $\eta = \eta(v) \in [\bar{u}_2, \hat{v}_2]$ such that*

$$f'(\eta) = \sigma(v, \eta). \tag{6.1}$$

The function $v \to \eta(v)$ is continuous, differentiable in $]\hat{v}_1, \bar{u}_2[$, and strictly decreasing, and it satisfies $\eta(\hat{v}_1) = \hat{v}_2$ and $\eta(\bar{u}_2) = \bar{u}_2$. Furthermore,

$$\sigma(v, \eta(v)) > \sigma(w, \eta(v)) \quad \forall\, v, w \;:\; \hat{v}_1 \leq v < w < \eta(v), \tag{6.2}$$

$$f'(v) > f'(\eta(v)) \quad \forall\, v \in \,]\hat{v}_1, \bar{u}_2[\,. \tag{6.3}$$

Proof. From assumptions (H5) and (H6), we deduce

$$f'(w) > f'(\bar{u}_2) \quad \forall w \in [\hat{v}_1, \hat{v}_2] \setminus \{\bar{u}_2\}. \tag{6.4}$$

Therefore,

$$\sigma(v, \bar{u}_2) = \frac{1}{\bar{u}_2 - v} \int_v^{\bar{u}_2} f'(w)dw > f'(\bar{u}_2) \quad \forall v \in [\hat{v}_1, \bar{u}_2[\,. \tag{6.5}$$

From (H5) and (H6), it also follows that there exists $\hat{v}_3 \in \,]\bar{u}_1, \bar{u}_2[$ such that

$$f'(\hat{v}_1) = f'(\hat{v}_2) = f'(\hat{v}_3), \tag{6.6}$$

$$\operatorname{sgn}[f'(v) - f'(\hat{v}_3)] = \operatorname{sgn}[\hat{v}_3 - v] \quad \forall\, v \in \,]\hat{v}_1, \hat{v}_2[\,. \tag{6.7}$$

The last equality implies that

$$\sigma(v, \hat{v}_2) < f'(\hat{v}_3) \quad \forall\, v \in [\hat{v}_3, \hat{v}_2[\,. \tag{6.8}$$

For any $v \in \,]\hat{v}_1, \hat{v}_3]$, we obtain by (H6), (6.6), and (6.7) that

$$\begin{aligned} \sigma(v, \hat{v}_2) &= \frac{1}{\hat{v}_2 - v} \left[(\hat{v}_2 - \hat{v}_1)\sigma(\hat{v}_1, \hat{v}_2) - \int_{\hat{v}_1}^{v} f'(w)dw \right] \\ &< \frac{1}{\hat{v}_2 - v} [(\hat{v}_2 - \hat{v}_1)f'(\hat{v}_3) - (v - \hat{v}_1)f'(\hat{v}_3)] = f'(\hat{v}_3). \end{aligned} \tag{6.9}$$

Inequalities (6.8), (6.9), and (6.6) yield

$$\sigma(v, \hat{v}_2) < f'(\hat{v}_2) \quad \forall\, v \in \,]\hat{v}_1, \hat{v}_2[\,. \tag{6.10}$$

From (6.5) and (6.10), we deduce that for any $v \in \,]\hat{v}_1, \bar{u}_2[$, there exists a value $\eta \in \,]\bar{u}_2, \hat{v}_2[$ such that (6.1) is satisfied. If $v = \hat{v}_1$ (resp. $v = \bar{u}_2$), then (6.1) holds with $\eta = \hat{v}_2$ (resp. $\eta = \bar{u}_2$). The uniqueness of η is checked analogously to the uniqueness of $\hat{v}$ in Lemma 4.1.

Let us now prove inequality (6.3). We observe that for any $v \in \,]\hat{v}_1, \bar{u}_2[\,$,

$$\hat{v}_3 < \bar{u}_2 < \eta(v) < \hat{v}_2. \tag{6.11}$$

Therefore, if $v \leq \hat{v}_3$, (6.3) follows from (6.7). Let us then consider the case where $v > \hat{v}_3$. If we define

$$M = \max\{f'(w) \,:\, w \in [v, \eta(v)]\},$$

we have

$$f'(\eta(v)) = \sigma(v, \eta(v)) = \frac{1}{\eta(v) - v} \int_v^{\eta(v)} f'(w)dw < M. \tag{6.12}$$

Since we are assuming $v > \hat{v}_3$, we deduce from (6.11) and (H5) that

$$M = \max\{f'(v), f'(\eta(v))\},$$

which implies (6.3) by virtue of (6.12).

Let us now define

$$H(v,w) = f'(w) - \sigma(v,w) \quad \text{for } (v,w) \in [\hat{v}_1, \bar{u}_2] \times [\bar{u}_2, \hat{v}_2].$$

Then $H(v,w) = 0$ if and only if $w = \eta(v)$. For any $v \in]\hat{v}_1, \bar{u}_2[$, we have by (6.1), (6.3) and (H5) that

$$H_v(v, \eta(v)) = \frac{f'(v) - \sigma(v, \eta(v))}{\eta(v) - v} > 0, \qquad H_w(v, \eta(v)) = f''(\eta(v)) > 0.$$

It follows that $v \to \eta(v)$ is differentiable and decreasing.

It remains to prove inequality (6.2). To this end, let us fix $v \in [\hat{v}_1, \bar{u}_2[$ and define

$$K(w) = f(w) - f(\eta(v)) - f'(\eta(v))(w - \eta(v)) \quad (w \in [v, \eta(v)]).$$

By (H5), (H6), (6.1), and (6.3), we have

$$K(v) = K(\eta(v)) = 0,$$

$$K'(v) = f'(v) - f'(\eta(v)) > 0 \quad \text{if } v > \hat{v}_1,$$

$$K'(v) = 0, \qquad K''(v) > 0, \quad \text{if } v = \hat{v}_1,$$

$$K'(\eta(v)) = 0 \qquad K''(\eta(v)) > 0.$$

It follows that the two endpoints $w = v$ and $w = \eta(v)$ are local minima for $K(\cdot)$. Suppose that $K(w) \leq 0$ for some $w \in]v, \eta(v)[$. Then there would exist at least three critical points for $K(\cdot)$ in $]v, \eta(v)[$. However, the derivative of K is $K'(w) = f'(w) - f'(\eta(v))$ and can vanish at at most one value of this interval by (H5) and (H6). The contradiction proves that $K(w)$ is positive for any $w \in]v, \eta(v)[$. Thus we conclude that

$$\sigma(v, \eta(v)) - \sigma(w, \eta(v)) = -\frac{K(w)}{w - \eta(v)} > 0$$

for any $w \in]v, \eta(v)[$. The proof is complete. □

Definition 6.2. *Let $t^* > 0$ be the value such that*

$$f'(W(\bar{u}_2, -t^*)) = f'(\bar{u}_2). \tag{6.13}$$

Let $v^* : [0, t^*] \to [\bar{u}_1, \bar{u}_2]$ be defined by $v^*(0) = \bar{u}_1$ and by the equality

$$f'(W(v^*(t), -t)) = f'(v^*(t)) \quad \forall\, t \in]0, t^*] \tag{6.14}$$

(see Remark 5.4). For $t > t^*$, let us set

$$v^*(t) = W(\bar{u}_2, t - t^*). \tag{6.15}$$

Given (x,t) with $t \geq 0$ and $x \in [f'(u_l)t, F(v^*(t), t)]$, let $\tilde{u}_1(x,t) \in [u_l, v^*(t)]$ be defined by

$$F(\tilde{u}_1(x,t), t) = x. \tag{6.16}$$

Similarly, given (x,t) *with* $t \geq 0$ *and* $x \in [F_+(\hat{v}_2,t), f'(u_r)t]$, *let* $\tilde{u}_2(x,t) \in [W(\hat{v}_2,t), u_r]$ *be defined by*

$$F(\tilde{u}_2(x,t),t) = x. \tag{6.17}$$

LEMMA 6.3. *The quantities introduced above are well defined. The function* v^* *is continuous, while* $\tilde{u}_1$ *and* $\tilde{u}_2$ *are classical solutions of equation* (1.1) *in the interior of their domain of definition and satisfy*

$$\tilde{u}_1(f'(u_l)t,t) = u_l, \tag{6.18}$$

$$\tilde{u}_2(F_+(\hat{v}_2,t),t) = W(\hat{v}_2,t), \qquad \tilde{u}_2(f'(u_r)t,t) = u_r. \tag{6.19}$$

Proof. By assumptions (H5) and (H7), there exists a unique value $w^* \in]u_l, \bar{u}_2[$ satisfying

$$f'(w^*) = f'(\bar{u}_2).$$

Since $W(\bar{u}_2, -t)$ decreases monotonically from $\bar{u}_2$ to u_l as t runs the interval $[0,\infty[$, there exists a unique value t^* for which

$$W(\bar{u}_2, -t^*) = w^*. \tag{6.20}$$

This is also the unique value which satisfies (6.13).

Following the proof of Lemma 5.2, we obtain that equality (6.14) defines v^* uniquely and that v^* is continuous at $t = 0$. From (6.13), (6.14), and (6.15), it follows that v^* is continuous at $t = t^*$ and that $v^*(t^*) = \bar{u}_2$.

To show that $\tilde{u}_1$ and $\tilde{u}_2$ are well defined, we need to prove

$$\partial_v F(v,t) > 0 \quad \forall\, t > 0, \quad \forall\, v \in]u_l, v^*(t)[\cup]W(\hat{v}_2,t), u_r[. \tag{6.21}$$

By (3.12) and (H3), this is equivalent to

$$f'(v) > f'(W(v,-t)) \quad \forall\, t > 0, \quad \forall\, v \in]u_l, v^*(t)[; \tag{6.22}$$

$$f'(W(v,t)) > f'(v) \quad \forall\, t > 0, \quad \forall\, v \in]\hat{v}_2, u_r[. \tag{6.23}$$

Inequality (6.23) follows from assumption (H5), while (6.22) follows from the definition of v^* for $t \leq t^*$. If $t > t^*$ then by (6.15) and (6.20),

$$W(v^*(t), -t) = W(\bar{u}_2, -t^*) = w^* \quad \forall\, t \geq t^*. \tag{6.24}$$

Let us now take $v \in]u_l, v^*(t)[$. If $v \leq \bar{u}_1$, then (6.22) is a consequence of assumption (H5). Otherwise, by (6.24) and (H5), we have

$$f'(v) \geq f'(\bar{u}_2) = f'(w^*) = f'(W(v^*(t), -t)) > f'(W(v,-t)).$$

Thus $\tilde{u}_1$ and $\tilde{u}_2$ are well defined. The other properties follow from the definition and from Lemma 3.4(v). □

LEMMA 6.4. *There exists a function* $\gamma \in C^1([0,T^*])$ *with* $T^* > t^*$ *which satisfies the following properties:*

$$\gamma(0) = 0, \qquad \gamma'(0) = f'(\hat{v}_1), \tag{6.25}$$

$$\gamma(t) \in \,]F(\hat{v}_1,t), F(v^*(t),t)[\quad \forall\, t \in \,]0,t^*], \tag{6.26}$$

$$\gamma(t) \in \,]F(\hat{v}_1,t), F(\bar{u}_2,t)[\quad \forall\, t \in [t^*,T^*[\,, \tag{6.27}$$

$$\gamma(T^*) = F(\bar{u}_2,T^*), \qquad \gamma'(T^*) = f'(\bar{u}_2), \tag{6.28}$$

$$\gamma'(t) = f'(\eta(\tilde{u}_1(\gamma(t),t))) \quad \forall\, t \in \,]0,T^*]. \tag{6.29}$$

Furthermore, if we set

$$\alpha(t) = \tilde{u}_1(\gamma(t),t), \qquad \omega(t) = \eta(\alpha(t)) \quad \textit{for } t \in \,]0,T^*], \tag{6.30}$$

we have

$$\alpha(t) \in \,]\hat{v}_1, \min\{\bar{u}_2, v^*(t)\}[\quad \forall\, t \in \,]0,T^*[\,. \tag{6.31}$$

$$\lim_{t\to 0} \alpha(t) = \hat{v}_1, \qquad \lim_{t\to 0} \omega(t) = \hat{v}_2, \tag{6.32}$$

$$\alpha(\cdot) \textit{ is increasing,} \qquad \omega(\cdot) \textit{ is decreasing.} \tag{6.33}$$

Remark 6.5. As in section 5, the curve γ we are dealing with will turn out to be a curve of discontinuity of the solution u of our problem. We will show that $u = \tilde{u}_1$ on the left of γ and that γ is a *right contact*, i.e., its speed is equal to the characteristic speed of u on the right. By Lemma 6.1, this means that the limit from the right of $u(\cdot,t)$ at a point $\gamma(t)$ is equal to $\eta(\tilde{u}_1(\gamma(t),t))$. Thus condition (3.2) yields equality (6.29). Observe that at time T^*, we have by (6.28) and the definition of $\tilde{u}$ that

$$u(\gamma(T^*)^-,T^*) = \tilde{u}_1(\gamma(T^*),T^*) = \bar{u}_2 = \eta(\bar{u}_2) = u(\gamma(T^*)^+,T^*).$$

Thus $u(\cdot,t)$ is no longer discontinuous across γ.

Proof of Lemma 6.4. To simplify notation, we set

$$\chi_1(t) = F(\hat{v}_1,t),$$

$$\chi_2(t) = \begin{cases} F(v^*(t),t) & \text{for } t \in [0,t^*], \\ F(\bar{u}_2,t) & \text{for } t \in [t^*,\infty[\,, \end{cases}$$

$$h(x,t) = f'(\eta(\tilde{u}_1(x,t)) \quad \text{for } t > 0,\ x \in [\chi_1(t),\chi_2(t)].$$

Then by assumptions (H5) and (H6), Definition 6.2, and Lemmas 3.4(iv) and 6.1, we have

$$h(\chi_1(t),t) = f'(\hat{v}_1) > f'(W(\hat{v}_1,-t)) = \chi_1'(t) \quad \forall\, t > 0, \tag{6.34}$$

$$h(\chi_2(t),t) = f'(\eta(v^*(t),t)) < f'(v^*(t)) = \chi_2'(t) \quad \forall\, t \in \,]0,t^*[\,, \tag{6.35}$$

$$h(x,t) < f'(\hat{v}_2) = f'(\hat{v}_1) \quad \forall t \in]0,t^*[\,,\ \forall x \in]\chi_1(t),\chi_2(t)[\,. \tag{6.36}$$

Following the proof of Lemma 5.6, we find $\gamma \in C^1([0,t^*])$ which satisfies properties (6.25), (6.26), (6.32), and (6.29) for $t \in]0,t^*]$. Then we continue with γ as a solution of (6.29) in the maximal interval $[t^*,T^*[$ where (6.27) is satisfied. If T^* is finite, we can define by continuity the value of γ at T^* because γ' is uniformly bounded by (6.29).

We can exclude that $T^* = \infty$. In fact, we have

$$\lim_{t\to\infty} \chi_2'(t) = \lim_{t\to\infty} f'(W(\bar{u}_2,-t)) = f'(u_l).$$

On the other hand, since η takes values in $[\bar{u}_2,\hat{v}_2]$, we have by (H5) and (H7) that

$$\gamma'(t) \geq f'(\bar{u}_2) > f'(u_l) \quad \forall\, t < T^*.$$

Thus inequality (6.27) cannot hold for arbitrarily large t, and T^* must be finite. By (6.34), we can also exclude that $\gamma(T^*) = \chi_1(T^*)$. Therefore equalities (6.28) are satisfied.

Inclusion (6.31) follows from (6.26) and (6.27), while (6.33) is proved by the same argument used for (5.20) in the proof of Lemma 5.6. □

We define γ after T^* by setting

$$\gamma(t) = \gamma(T^*) + F_+(\bar{u}_2, t-T^*) \quad \text{for } t > T^*. \tag{6.37}$$

Then by (6.28) and Definition 3.3,

$$\gamma(t) = F(\bar{u}_2,T^*) + F_+(\bar{u}_2,t-T^*) = F(W(\bar{u}_2,t-T^*),t),$$

and by (6.16), we obtain

$$\tilde{u}_1(\gamma(t),t) = W(\bar{u}_2,t-T^*) \quad \forall\, t > T^*. \tag{6.38}$$

The next result states that the region lying between the curves $x = \gamma(t)$ and $x = F_+(\hat{v}_2,t)$ is covered univalently by the characteristics emanating from γ. Therefore, it is possible to define in this region a function $\hat{u}$ which assumes the desired values along γ and solves equation (1.1).

Lemma 6.6. *For any (x,t) with $t > 0$ and $x \in [\gamma(t),F_+(\hat{v}_2,t)]$, there exists a unique value $\tau = \tau(x,t) \in [0,\min\{t,T^*\}]$ such that*

$$x = \gamma(\tau) + F_+(\omega(\tau),t-\tau). \tag{6.39}$$

Proof. For fixed $t > 0$, define

$$y(\tau) = \gamma(\tau) + F_+(\omega(\tau),t-\tau) \quad \text{for } \tau \in [0,\min\{t,T^*\}].$$

Then by (6.32) and (6.37),

$$y(0) = F_+(\hat{v}_2,t), \qquad y(\min\{t,T^*\}) = \gamma(t).$$

We only need to prove that $y(\cdot)$ is strictly decreasing. Since f is convex in $[\bar{u}_2,u_r]$, we have $\partial_v F_+(v,s) > 0$ for any $v \in [\bar{u}_2,u_r]$ and $s > 0$. Furthermore, by (6.33), $\omega' \leq 0$. Thus by (6.29) and (3.5), we conclude

$$\begin{aligned} y'(\tau) &= \gamma'(\tau) + \partial_v F_+(\omega(\tau),t-\tau)\,\omega'(\tau) - \partial_t F_+(\omega(\tau),t-\tau) \\ &\leq f'(\omega(\tau)) - f'(W(\omega(\tau),t-\tau)) < 0. \qquad \square \end{aligned}$$

Definition 6.7. *For any $t > 0$ and $x \in [\gamma(t), F_+(\hat{v}_2, t)]$, set*

$$\hat{u}(x,t) = W(\omega(\tau(x,t)), t - \tau(x,t)), \tag{6.40}$$

where ω and τ are defined as in Lemmas 6.4 *and* 6.6.

Lemma 6.8. *The function $\hat{u}$ introduced above is a classical solution of equation* (1.1) *in the interior of its domain of definition. Moreover, it satisfies*

$$\hat{u}(\gamma(t), t) = \begin{cases} \omega(t) & \text{if } t \in [0, T^*], \\ W(\bar{u}_2, t - T^*) & \text{if } t \in [T^*, \infty[\,, \end{cases} \tag{6.41}$$

$$\hat{u}(F_+(\hat{v}_2, t)) = W(\hat{v}_2, t). \tag{6.42}$$

Proof. By (6.39), (6.40), and (3.10), we have

$$\begin{aligned} x - \gamma(\tau(x,t)) &= F_+(\omega(\tau(x,t)), t - \tau(x,t)) \\ &= F(W(\omega(\tau(x,t)), t - \tau(x,t)), t - \tau(x,t)) \\ &= F(\hat{u}(x,t), t - \tau(x,t)). \end{aligned} \tag{6.43}$$

Furthermore, by (6.29), (3.11), and (6.40),

$$\begin{aligned} \gamma'(\tau(x,t)) &= f'(\omega(\tau(x,t))) \\ &= f'(W(\hat{u}(x,t), \tau(x,t) - t)) \\ &= \partial_t F(\hat{u}(x,t), t - \tau(x,t)). \end{aligned}$$

Thus if we differentiate (6.43), we obtain

$$1 = \partial_v F \, \partial_x \hat{u}, \qquad 0 = \partial_v F \, \partial_t \hat{u} + \partial_t F.$$

Following the proof of Lemma 3.4(v), we can conclude that $\hat{u}$ solves equation (1.1). Equalities (6.41) and (6.42) follow from the definition of $\hat{u}$. □

Theorem 6.9. *Let assumptions* (H1), (H3), *and* (H5)–(H7) *be satisfied and let $\tilde{u}_1$, $\tilde{u}_2$, and $\hat{u}$ be defined as in Definitions* 6.2 *and* 6.7. *Then the function γ given by Lemma* 6.4 *is unique, and the solution of problem* (1.1)–(1.2) *is*

$$u(x,t) = \begin{cases} u_l, & x < f'(u_l)t, \\ \tilde{u}_1(x,t), & f'(u_l)t < x < \gamma(t), \\ \hat{u}(x,t), & \gamma(t) < x < F_+(\hat{v}_2, t), \\ \tilde{u}_2(x,t), & F_+(\hat{v}_2, t) < x < f'(u_r)t, \\ u_r, & x > f'(u_r)t. \end{cases} \tag{6.44}$$

Furthermore, $u(\cdot, t)$ is discontinuous at $\gamma(t)$ for $t < T^*$, while it is continuous everywhere for $t \geq T^*$ (T^* has been introduced in Lemma 6.4).

Proof. The function u is a classical solution of equation (1.1) in each of the five regions defined above by virtue of assumption (H3) and Lemmas 6.3 and 6.8. By equalities (6.18), (6.19), and (6.42), we deduce that u is continuous along the curves $x = f'(u_l)t$, $x = f'(u_r)t$, and $x = F_+(\hat{v}_2, t)$, while (6.38) and (6.41) imply the continuity along γ for $t \geq T^*$. By Lemmas 6.1, 6.4, and 6.8, γ is a curve of discontinuity which satisfies conditions (3.2) and (3.3) for $t < T^*$. Finally, the uniqueness of γ follows from the uniqueness of u. $\square$

From the previous theorem, it is possible to deduce the following asymptotic representation for u. We omit the proof, which is similar to the proof of Theorem 2.3.

THEOREM 6.10. *Under assumptions* (H1), (H3), *and* (H5)–(H7), *the solution of problem* (1.1)–(1.2) *satisfies*

$$u(x,t) = \begin{cases} u_l & \text{for } x < f'(u_l)t, \\ \\ \phi(x - f'(u_l)t) + o(1) & \text{for } x > f'(u_l)t, \end{cases} \tag{6.45}$$

where $\phi : [0, \infty[\to [u_l, u_r[$ *is defined by*

$$\int_{u_l}^{\phi(\xi)} \frac{f'(v) - f'(u_l)}{g(v)}\, dv = \xi \quad \text{for any } \xi \geq 0.$$

Acknowledgments. The author wishes to thank the referees for their helpful suggestions.

REFERENCES

[1] D. P. BALLOU, *Solutions to nonlinear hyperbolic Cauchy problems without convexity conditions*, Trans. Amer. Math. Soc., 152 (1970), pp. 441–460.

[2] L. L. BONILLA, *Solitary waves in semiconductors with finite geometry and the Gunn effect*, SIAM J. Appl. Math., 51 (1991), pp. 727–747.

[3] C. M. DAFERMOS, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[4] C. M. DAFERMOS, *Large time behaviour of solutions of hyperbolic balance laws*, Bull. Greek Math. Soc., 25 (1984), pp. 15–29.

[5] C. M. DAFERMOS, *Regularity and large time behavior of solutions of a conservation law without convexity*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1985), pp. 201–239.

[6] H. FAN AND J. K. HALE, *Large-time behaviour in inhomogeneous conservation laws*, Arch. Rational Mech. Anal., 125 (1993), pp. 201–216.

[7] H. FAN AND J. K. HALE, *Attractors in inhomogeneous conservation laws and parabolic regularizations*, Trans. Amer. Math. Soc., 347 (1995), pp. 1239–1254.

[8] E. GODLEWSKI AND P. RAVIART, *Hyperbolic Systems of Conservation Laws*, Mathématiques et Applicationes, Vols. 3/4, Société de Mathématiques Appliquées et Industrielles, Ellipses, Paris, 1991.

[9] B. W. KNIGHT AND G. A. PETERSON, *Nonlinear analysis of the Gunn effect*, Phys. Rev., 147 (1966), pp. 617–621.

[10] S. N. KRUZHKOV, *First order quasilinear equations in several independent variables*, Mat. Sb., 81 (1970), pp. 228–255 (in Russian); Math. USSR-Sb., 10 (1970), pp. 217–243 (in English).

[11] S. N. KRUZHKOV AND N. S. PETROSJAN, *Asymptotic behaviour of the solutions of the Cauchy problem for nonlinear first order equations*, Uspekhi Mat. Nauk, 42 (1987), pp. 3–40 (in Russian); Russian Math. Surveys, 42 (1987), pp. 1–47 (in English).

[12] C. Z. LI AND T. P. LIU, *Asymptotic states for hyperbolic conservation laws with a moving source*, Adv. Appl. Math., 4 (1983), pp. 353–379.

[13] T. P. LIU, *Nonlinear resonance for quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.

[14] A. N. LYBEROPOULOS, *Asymptotic oscillations of solutions of scalar conservation laws with convexity under the action of a linear excitation*, Quart. Appl. Math., 48 (1990), pp. 755–765.

[15] A. N. LYBEROPOULOS, *Large time structure of solutions of scalar conservation laws without convexity in the presence of a linear source field*, J. Differential Equations, 99 (1992), pp. 342–380.

[16] A. N. Lyberopoulos, *A Poincaré–Bendixson theorem for scalar balance laws*, Proc. Roy. Soc. Edinburgh Sect A, 124 (1994), pp. 589–607.

[17] J. D. Murray, *On the Gunn effect and other physical examples of perturbed conservation equations*, J. Fluid Mech., 44 (1970), pp. 315–346.

[18] R. Natalini and A. Tesei, *On a class of perturbed conservation laws*, Adv. Appl. Math., 13 (1992), pp. 429–453.

[19] O. A. Oleinik, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi–linear equation*, Uspekhi Mat. Nauk, 14 (1959), pp. 165–170 (in Russian); Amer. Math. Soc. Transl. Ser. 2, 33 (1963), pp. 285–290 (in English).

[20] C. Sinestrari, *Large time behaviour of solutions of balance laws with periodic initial data*, Nonlinear Differential Equations Appl., 2 (1995), pp. 111–131.

[21] C. Sinestrari, *Asymptotic profile of solutions of conservation laws with source*, Differential Integral Equations, 9 (1996), pp. 499–525.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 136–161, January 1997

ON THE RATE OF CONVERGENCE TO EQUILIBRIUM FOR A SYSTEM OF CONSERVATION LAWS WITH A RELAXATION TERM*

ASLAK TVEITO[†] AND RAGNAR WINTHER[†]

Abstract. We analyze a simple system of conservation laws with a strong relaxation term. Well-posedness of the Cauchy problem in the framework of bounded-total-variation (BV) solutions is proved. Furthermore, we prove that the solutions converge towards the solution of an equilibrium model as the relaxation time $\delta > 0$ tends to zero. Finally, we show that the difference between an equilibrium solution ($\delta = 0$) and a nonequilibrium solution ($\delta > 0$) measured in L^1 is bounded by $O(\delta^{1/3})$.

Key words. hyperbolic conservation laws, relaxation terms, nonequilibrium, rate of convergence

AMS subject classifications. 35L65, 65M10.

PII. S0036141094263755

1. Introduction.

The purpose of this paper is to study the following system of conservation laws:

$$(u+v)_t + f(u)_x = 0, \qquad (1.1)$$
$$\delta v_t = A(u) - v.$$

Here f and A are given functions, u and v are the unknowns and $\delta > 0$ is referred to as the relaxation time. The function $A(u)$ will throughout this paper be assumed to be an increasing function of u. Further assumptions on the model will be given in section 2.

We will be concerned in particular with the convergence of $(u, v) = (u_\delta, v_\delta)$ as δ tends to zero. In the limit of zero relaxation time, a scalar conservation law of the form

$$(w + A(w))_t + f(w)_x = 0 \qquad (1.2)$$

is obtained. Systems of the form (1.1) are usually referred to as "nonequilibrium" models, whereas (1.2) is the "equilibrium" model. The main result of this paper is that for proper conditions on the initial data, the solutions of the nonequilibrium model tend to the solution of the equilibrium model in L^1 with a deviation bounded by $O(\delta^{1/3})$.

System (1.1) arises in chromatography and is discussed in [19, 20, 28]. In this framework, u denotes the density of some species contained in a fluid flowing through a fixed bed and v denotes the density of the species adsorbed on the material in the bed. The right-hand side of the second equation models the adsorption. Different forms of adsorption functions A are discussed by, e.g., Bear and Bachmat [1, Chapter 6].

The equilibrium assumption in chromatography states that the rate of the chemical reaction is so large that the reaction can be considered as instantaneous compared to the time scales of other effects. We prove that for initial data close to equilibrium

* Received by the editors February 25, 1995; accepted for publication (in revised form) August 29, 1995. This research was supported by Norwegian Research Council (NFR) program STP.29643 at the Department of Applied Mathematics, Sintef, Oslo, Norway.

http://www.siam.org/journals/sima/28-1/26375.html

[†] Department of Informatics, P. O. Box 1080, Blindern, University of Oslo, Norway (aslak@ifi.uio.no, ragnar@ifi.uio.no).

and for small relaxation times δ, the equilibrium model provides good approximations of the nonequilibrium solutions. One should, however, be cautious. In [24], the system in a radial geometry is considered as a model for near-well reservoir simulation. Their results indicate the equilibrium model is inadequate for the analysis of this problem.

A nice introduction to systems of conservation laws with relaxation terms can be found in Whitham's book [28], where such models arising in, e.g., chromatography, traffic modeling, water waves, and gas dynamics are discussed. For applications in chromatography, a detailed discussion is provided in the books of Rhee, Aris, and Amundson [19, 20]. Systems consisting of one conservation equation and one equation with a relaxation term have been studied by several authors; cf. [2, 5, 6, 7, 12, 13, 15, 17, 18, 21, 24, 25, 27]. Much of this research is motivated by combustion theory and especially by Majda's model [15]. The question of well-posedness of this system is discussed by Teng and Ying [25] and Levy [12], and computational studies of the system are presented by Colella, Majda, and Roytburd [5] and Pember [17, 18].

The so-called subcharacteristic condition plays a central role in systems with relaxation terms. Consider system (1.1) above and let $\lambda_1 = 0$ and $\lambda_2 = f'$ denote the characteristic speeds. Similarly, $\lambda^* = f'/(1 + A')$ denotes the characteristic speed of (1.2). Then the subcharacteristic condition states that $\lambda_1 \leq \lambda^* \leq \lambda_2$, which due to the monotonicity of A is satisfied for our models. It turns out that the subcharacteristic condition is necessary for the stability of the nonequilibrium model. This issue is analyzed using linearization by Whitham [28, Chapter 10], for "small waves" by Liu [13], for a linear model by LeVeque and Wang [11], and finally by Chen and Liu [2]. In the latter paper, Chen and Liu actually prove convergence of solutions of the nonequilibrium towards the solution of an equilibrium model for two different systems. Their results are based on the theory of compensated compactness. Generalizations of their results are given in the recent paper by Chen, Levermore, and Liu [3]. Also, Schochet [21] proves convergence of a family of nonequilibrium solutions to the solution of an equilibrium model. He studies the refined traffic model, introduced by Whitham [28], which consists of a standard conservation law for the density of cars and a nonequilibrium model for the velocity.

The plan of this paper is as follows. We begin in section 2 by giving precise assumptions on system (1.1) and the initial data. Furthermore, we state the main results of the paper. Section 3 is devoted to the analysis of a finite-difference scheme approximating system (1.1). The results of section 3 are used in section 4, where we demonstrate existence, uniqueness, and stability of solutions of (1.1). We should mention here that the well-posedness of (1.1) was proved by both Levy [12] and Teng and Ying [25]. Their estimates, however, depend on the relaxation time, and since we will study the convergence of solutions of (1.1) as $\delta \to 0$, we need δ independent estimates.

In section 5, we introduce an auxiliary system where small diffusion terms are present. This auxiliary system enables us to prove in section 6 that the solution of (1.1) tends to the solution of (1.2) in L^1 with a deviation bounded by $O(\delta^{1/3})$.

Remarks. (1) The results obtained in this paper could without too much difficulty be generalized to yield slightly more general systems than those covered by (1.1). For clarity, however, we have chosen to analyze a model that is as simple as possible in order to avoid messy details containing no essential new insights.

(2) We do not know if the convergence estimate of the form $O(\delta^{1/3})$ is optimal. However, computational experiments that we have done seem to indicate that if an estimate of the form $O(\delta^{\gamma})$ is sought, then the optimal value of γ is somewhere in the

interval $[1/4, 1/2]$; see also section 6 of [22].

2. Preliminaries and statement of the main results. The purpose of this paper is to study how well solutions of the scalar conservation law

$$(w + A(w))_t + f(w)_x = 0 \tag{2.1}$$

approximate the corresponding solutions of the system

$$\begin{aligned} (u+v)_t + f(u)_x &= 0, \\ \delta v_t &= A(u) - v \end{aligned} \tag{2.2}$$

for small positive values of the relaxation time δ, i.e., $\delta \in (0, 1]$. The explicit dependence of the solution (u, v) on δ will usually be suppressed. We will assume that $f = f(u)$ is a smooth function satisfying $f(0) = 0$ and that

$$f'(u) \geq 0 \quad \text{for } u \in [0, 1].$$

The function $A = A(u)$ will be assumed to satisfy the requirements

$$\begin{gathered} A(0) = 0, \qquad A(1) = 1, \\ A'(u) \geq 0, \\ |A''(u)| \leq \alpha < 1 \end{gathered} \tag{2.3}$$

for all relevant values of u. We shall consider solutions of (2.2) in the state space

$$\mathcal{S} = [0, 1] \times [0, 1]$$

and solutions of (2.1) in $[0, 1]$.

In equilibrium, we have $v = A(u)$, and system (2.2) degenerates to the scalar equation (2.1). In order to study the deviation from equilibrium in the nonequilibrium model, we introduce an auxiliary variable defined by

$$p = A(u) - v.$$

The initial conditions (u^0, v^0) for system (2.2) are supposed to satisfy the following:

$$\begin{array}{rc} \text{(i)} & (u^0(x), v^0(x)) \in \mathcal{S} \quad \forall x \in \mathcal{R}, \\ \text{(ii)} & \mathrm{TV}(u^0) + \mathrm{TV}(v^0) \leq M, \\ \text{(iii)} & \|p^0\|_1 = \|A(u^0) - v^0\|_1 \leq M\delta, \\ \text{(iv)} & u^0(\pm\infty), v^0(\pm\infty) = 0. \end{array} \tag{2.4}$$

Here and in the rest of this paper, M denotes a generic finite constant independent of δ. The L^1-norm is denoted by $\|\cdot\|_1$ and $\mathrm{TV}(\cdot)$ denotes the total variation, defined by

$$\mathrm{TV}(z) = \sup_{h \neq 0} \int_{\mathcal{R}} \frac{|z(x+h) - z(x)|}{|h|} dx.$$

Furthermore, $\mathrm{BV} = \mathrm{BV}(\mathcal{R})$ denotes the subspace of L^1_{loc} consisting of functions with bounded total variation.

Note that (iii) above assures that the initial conditions are close to equilibrium. This is a natural assumption since we primarily want to analyze the convergence towards the solutions of the equilibrium model in (2.1). However, we shall also indicate that if (iii) is not satisfied initially, an initial layer will appear and $\|p(t)\|_1 = O(\delta)$ will be reached at an exponential rate.

What we mean by an entropy solution of the scalar equation (2.1) is well known, but let us define a similar concept for (2.2). For any $T > 0$, we let $\mathcal{D}_+(T)$ denote the set of all nonnegative C^∞-functions with compact support in $\mathcal{R} \times [0, T]$.

DEFINITION 1. *Let* (u^0, v^0) *satisfying* (2.4) *be the given initial data. Then* (u, v) *is called an* entropy solution of system (2.2) *if the following requirements are satisfied:*

1. $(u, v) \in \mathcal{S}\ \forall (x, t) \in \mathcal{R} \times \mathcal{R}_0^+$;
2. $\mathrm{TV}(u(\cdot, t)), \mathrm{TV}(v(\cdot, t)) \le M\ \forall t \in \mathcal{R}_0^+$;
3. $\|u(\cdot, t) - u(\cdot, \tau)\|_1 + \|v(\cdot, t) - v(\cdot, \tau)\|_1 \le M|t - \tau|\ \forall t, \tau \in \mathcal{R}_0^+$;
4. *For any* $(k, q) \in \mathcal{S}$ *and any* $\varphi, \psi \in \mathcal{D}_+(T)$*, where* $T > 0$ *is arbitrary,*

$$\begin{aligned}
&\int_0^T \int_{\mathcal{R}} [|u - k|\varphi_t + |f(u) - f(k)|\varphi_x + |v - q|\psi_t] dx\, dt \\
&\quad + \int_{\mathcal{R}} [|u^0 - k|\varphi(x, 0) + |v^0 - q|\psi(x, 0)] dx \\
&\quad - \int_{\mathcal{R}} [|u(x, T) - k|\varphi(x, T) + |v(x, T) - q|\psi(x, T)] dx \\
&\quad \ge \frac{1}{\delta} \int_0^T \int_{\mathcal{R}} [\sigma(u - k)\varphi - \sigma(v - q)\psi](A(u) - v) dx\, dt,
\end{aligned} \tag{2.5}$$

where σ *denotes the sign function.*

Remark. Let us again mention that the existence of entropy solutions of systems of the form of (2.2) has previously been proved by Levy [12] and Teng and Ying [25]. The reason for us to redo the existence part here is that we shall need δ-independent estimates when we consider the convergence of (u_δ, v_δ) as $\delta \to 0$. Such estimates are not provided by [12, 25].

Existence of an entropy solution satisfying the requirements of Definition 1 will be proved using a finite-difference scheme. We note that existence could also be proved using the parabolic regularizations discussed below, but the problem of computing solutions of systems of the form of (2.2) is important, and hence we find it desirable to prove existence by demonstrating convergence of a numerical scheme. An error estimate for the scheme is derived in [22].

Our scheme for analyzing system (1.1) is straightforward and semiimplicit:

$$\begin{aligned}
\frac{(u_j^{n+1} + v_j^{n+1}) - (u_j^n + v_j^n)}{\Delta t} + \frac{f(u_j^n) - f(u_{j-1}^n)}{\Delta x} &= 0 \\
\frac{\delta(v_j^{n+1} - v_j^n)}{\Delta t} &= A(u_j^{n+1}) - v_j^{n+1}
\end{aligned} \tag{2.6}$$

Here u_j^n denotes an approximation of $u(x, t)$ over the grid block

$$B_j^n = [x_{j-1/2}, x_{j+1/2}) \times [t_n, t_{n+1}),$$

where $x_j = j\Delta x$ and $t_n = n\Delta t$. Furthermore, Δt and Δx denote the step lengths in the t and x directions, respectively. Similarly, v_j^n approximates v on B_j^n. The scheme is initialized by setting

$$u_j^0 = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u^0(x)\,dx \quad \text{and} \quad v_j^0 = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} v^0(x)\,dx. \tag{2.7}$$

Throughout the paper, we will assume that the grid parameters satisfy the Courant–Friedrichs–Lewy (CFL) condition

$$\mu f'(u) \le 1 \tag{2.8}$$

for all $u \in [0,1]$, where $\mu = \Delta t/\Delta x$. It is interesting from a computational point of view to note that this condition is independent of the relaxation time δ; thus very small relaxation times do not force us to use equally small time steps.

For a grid function u, the discrete total variation is defined by

$$\mathrm{TV}(u) = \sum_{j\in\mathcal{Z}} |u_j - u_{j-1}|,$$

and $\|\cdot\|_1$ denotes the discrete L^1-norm defined by

$$\|u\|_1 = \Delta x \sum_{j\in\mathcal{Z}} |u_j|.$$

In order to study the deviation from equilibrium in the nonequilibrium system, we define

$$p_j^n = A(u_j^n) - v_j^n.$$

We consider initial data satisfying

$$\begin{array}{rc} \text{(i)} & (u_j^0, v_j^0) \in S \quad \text{for all } j \in \mathcal{Z}, \\ \text{(ii)} & \mathrm{TV}(u^0) + \mathrm{TV}(v^0) \le M, \\ \text{(iii)} & \|p^0\|_1 \le M\delta, \\ \text{(iv)} & (u_{-\infty}^0, v_{-\infty}^0) = (u_\infty^0, v_\infty^0) = 0, \end{array} \tag{2.9}$$

where M is a finite constant independent of the grid parameters and the relaxation time. Note that (2.9) is simply a discrete version of (2.4).

Based on the results of the scheme, a family of approximate solutions is defined by setting

$$(u_\Delta, v_\Delta)(x,t) = (u_j^n, v_j^n) \qquad \text{for } (x,t) \in B_j^n \quad \forall (j,n) \in \mathcal{Z} \times \mathcal{Z}^+.$$

By using the properties of the scheme presented above, the following theorem will be proved in section 4.

THEOREM 2.1. *Let (u^0, v^0) be initial data satisfying* (2.4) *and let (u_j^0, v_j^0) be the corresponding discrete initial values generated by* (2.7)*. Then as Δx and Δt tend to zero, the family $\{(u_\Delta, v_\Delta)\}$ of approximate solutions generated by the finite-difference scheme described above converges in $(L^1_{\mathrm{loc}}(\mathcal{R}\times\mathcal{R}_0^+))^2$ towards a pair of functions (u,v). Furthermore, the limit (u,v) is a unique entropy solution satisfying the requirements of Definition* 1.

The entropy solution (u,v) satisfies

$$\|p(\cdot,t)\|_1 = \|A(u) - v\|_1 \le M\delta \quad \text{for} \quad t \ge 0,$$

where M is a finite constant not depending on δ. If $(\bar{u}^0, \bar{v}^0)$ is another pair of initial data satisfying (2.4)*, there is a unique entropy solution $(\bar{u}, \bar{v})$ such that*

$$\|u(\cdot,t) - \bar{u}(\cdot,t)\|_1 + \|v(\cdot,t) - \bar{v}(\cdot,t)\|_1 \leq \|u^0 - \bar{u}^0\|_1 + \|v^0 - \bar{v}^0\|_1 \quad \text{for} \;\; t \geq 0. \qquad \square$$

Since our existence theorem is based on estimates independent of δ, we could by appealing to Helly's theorem prove convergence of a subsequence in L^1_{loc} of $(u,v) = (u_\delta, v_\delta)$ as $\delta \to 0$. However, we want more than convergence—we are also interested in the rate of convergence. In order to analyze this issue, we are lead to study a parabolic regularization of the system

$$\begin{aligned} u^\epsilon_t + f(u^\epsilon)_x &= \frac{1}{\delta}(v^\epsilon - A(u^\epsilon)) + \epsilon u^\epsilon_{xx}, \\ v^\epsilon_t &= \frac{1}{\delta}(A(u^\epsilon) - v^\epsilon) + \epsilon v^\epsilon_{xx}. \end{aligned} \tag{2.10}$$

The properties of this system are analyzed in section 5. In section 6, we begin by proving that (cf. Lemma 6.1)

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_1 + \|v^\epsilon(\cdot,t) - v(\cdot,t)\|_1 \leq M\epsilon^{1/2}, \tag{2.11}$$

where (u^ϵ, v^ϵ) solves the regularized problem and (u,v) solves (2.2). Then we introduce a regularized equilibrium model of the form

$$(w^\epsilon + A(w^\epsilon))_t + f(w^\epsilon)_x = \epsilon(w^\epsilon + A(w^\epsilon))_{xx}. \tag{2.12}$$

For this scalar equation, it is well known (cf. Kuznetsov [10]) that

$$\|w - w^\epsilon\|_1 \leq M\epsilon^{1/2}, \tag{2.13}$$

where w solves (2.1) and w^ϵ solves (2.12). Furthermore, we prove (cf. Lemma 6.2) that

$$\|u^\epsilon - w^\epsilon\|_1 \leq \frac{M\delta}{\epsilon}, \tag{2.14}$$

where δ is the relaxation time and ϵ is the diffusion coefficient. Collecting these results, we observe that

$$\|u - w\|_1 \leq \|u - u^\epsilon\|_1 + \|u^\epsilon - w^\epsilon\|_1 + \|w^\epsilon - w\|_1 \leq M\left(\epsilon^{1/2} + \frac{\delta}{\epsilon} + \epsilon^{1/2}\right).$$

Hence, by choosing $\epsilon = \delta^{2/3}$, we obtain the following result.

THEOREM 2.2. *Let (u^0, v^0) be a pair of initial data satisfying* (2.4)*, and let $w^0 = u^0$. Then for any finite $T > 0$, there is a finite constant M such that*

$$\|u(\cdot,t) - w(\cdot,t)\|_1 \leq M\delta^{1/3} \quad \textit{for all} \;\; 0 \leq t \leq T.$$

Here (u,v) and w are solutions of (2.2) *and* (2.1)*, respectively.* □

The rest of this paper is devoted to the proof of Theorems 2.1 and 2.2.

Remark. Heuristic arguments, e.g., by a Chapman–Enskog expansion, indicate a rate of 1/2 rather than 1/3 in Theorem 2.2, but we have not been able to give rigorous arguments for such an improved rate. Neither have we been able to prove that 1/3 is optimal. In estimates (2.11), (2.13), and (2.14) leading to this rate, the first and second results are sharp but estimate (2.14) is probably not sharp. Our attempts at improving this estimate have not been successful.

3. Bounds on the approximate solutions. In this section, we will study the properties of the finite-difference scheme in (2.6) approximating the solutions of (2.2). In fact, the existence of an entropy solution of the system is proved by demonstrating convergence of the family of approximate solutions generated by the finite-difference scheme. We start by considering the standard estimates needed to prove the existence of a bounded-total-variation (BV) solution: an L^∞-bound, a bound on the total variation, and finally an L^1-continuity estimate in time. With these bounds, independent of the mesh size and the relaxation time δ, the existence of a weak solution can be demonstrated; cf., e.g., Smoller [23].

We also show that if the initial data is close to equilibrium, the discrete solution remains close to equilibrium for all time. Finally, we show that the finite difference solutions satisfy a discrete entropy inequality. This latter property enables us to prove the existence of an entropy solution satisfying the requirements of Definition 1.

The finite-difference approximations have the following properties.

LEMMA 3.1. *Suppose that the initial data* (u^0, v^0) *satisfy* (2.9) *and that the grid parameters* Δt *and* Δx *satisfy the CFL condition* (2.8). *Then for any* $\delta > 0$, *there is a finite constant* M *independent of* δ, Δt, *and* Δx *such that*

$$\begin{aligned}
&\text{I.} && (u_j^n, v_j^n) \in \mathcal{S} \quad \textit{for all } (j,n) \in \mathcal{Z} \times \mathcal{Z}^+,\\
&\text{II.} && \mathrm{TV}(u^n) + \mathrm{TV}(v^n) \le \mathrm{TV}(u^0) + \mathrm{TV}(v^0),\\
&\text{III.} && \|p^n\|_1 \le M\delta,\\
&\text{IV.} && \|u^n - u^m\|_1 + \|v^n - v^m\|_1 \le M|n-m|\Delta t,\\
&\text{V.} && \|u^n - \bar{u}^n\|_1 + \|v^n - \bar{v}^n\|_1 \le \|u^0 - \bar{u}^0\|_1 + \|v^0 - \bar{v}^0\|_1,
\end{aligned}$$

where $(\bar{u}^n, \bar{v}^n)$ *is a discrete solution based on initial data* $(\bar{u}^0, \bar{v}^0)$ *satisfying* (2.9). □

Note that since the entropy solution will inherit the properties of the approximate solutions, it will satisfy a maximum principle (I), be TV stable (II), be close to equilibrium (III), be L^1-continuous in time (IV), and be stable in L^1 (V).

The next subsections are devoted to the proof of Lemma 3.1.

3.1. Proof of I: Maximum principle. For a fixed pair of (j,n), we let $(u,v) = (u_j^{n+1}, v_j^{n+1})$, $(\bar{u}, \bar{v}) = (u_j^n, v_j^n)$, and $(u^L, v^L) = (u_{j-1}^n, v_{j-1}^n)$. Then by (2.6), we get the equations

$$\begin{aligned}
u + v &= \bar{u} + \bar{v} - \frac{\Delta t}{\Delta x}(f(\bar{u}) - f(u^L),\\
(\delta + \Delta t)v &- \Delta t A(u) = \delta \bar{v}.
\end{aligned} \tag{3.1}$$

Let us first verify that (3.1) has a unique solution (u,v) for any given $(\bar{u}, \bar{v}, u^L, v^L)$. Denote by r_1 and r_2 the two right-hand sides such that

$$v = r_1 - u,$$

and thus the single equation

$$K(u) \equiv \Delta t A(u) + (\delta + \Delta t)u + r_2 - r_1(\delta + \Delta t) = 0$$

determines u. Since K is monotone and $K(\pm\infty) = \pm\infty$, it is clear that (u,v) is uniquely determined by (3.1).

Next, we assume that $0 \le \bar{u}, \bar{v}, u^L, v^L \le 1$, and we want to prove that this implies

$$0 \le u, v \le 1, \tag{3.2}$$

which, in fact, proves part I of Lemma 3.1. In order to prove (3.2), we consider u and v as functions of u^L, $\bar{u}$, and $\bar{v}$, i.e.,

$$u = u(u^L, \bar{u}, \bar{v}) \quad \text{and} \quad v = u(u^L, \bar{u}, \bar{v}).$$

Now we want to show that u and v are monotone in all arguments, and we start by considering $\frac{\partial u}{\partial \bar{v}}$ and $\frac{\partial v}{\partial \bar{v}}$. From the first equation of (3.1), we get

$$\frac{\partial u}{\partial \bar{v}} + \frac{\partial v}{\partial \bar{v}} = 1, \tag{3.3}$$

and the second equation gives

$$(\delta + \Delta t)\frac{\partial v}{\partial \bar{v}} - \Delta t A'(u)\frac{\partial u}{\partial \bar{v}} = \delta;$$

hence

$$\frac{\partial v}{\partial \bar{v}} = \frac{\delta + \Delta t A'(u)}{\delta + \Delta t + \Delta t A'(u)} \in (0,1),$$

and then by (3.3), $\frac{\partial u}{\partial \bar{v}} \in (0,1)$ and we conclude that

$$\frac{\partial u}{\partial \bar{v}}, \frac{\partial v}{\partial \bar{v}} > 0. \tag{3.4}$$

Similarly, by differentiating (3.1) with respect to $\bar{u}$, we get

$$\frac{\partial u}{\partial \bar{u}} + \frac{\partial v}{\partial \bar{u}} = 1 - \mu f'(\bar{u}) \tag{3.5}$$

and

$$(\delta + \Delta t)\frac{\partial v}{\partial \bar{u}} - \Delta t A'(u)\frac{\partial u}{\partial \bar{u}} = 0;$$

hence

$$\frac{\partial v}{\partial \bar{u}} = \frac{\Delta t A'(u)}{\delta + \Delta t + \Delta t A'(u)}(1 - \mu f'(\bar{u})) \equiv \omega(1 - \mu f'(\bar{u})), \tag{3.6}$$

where $\omega \in [0,1)$. Now (3.5) gives

$$\frac{\partial u}{\partial \bar{u}} = (1 - \mu f'(\bar{u}))(1 - \omega),$$

and then by the CFL condition (2.8), we can conclude that

$$\frac{\partial u}{\partial \bar{u}}, \frac{\partial v}{\partial \bar{u}} \geq 0. \tag{3.7}$$

Finally, by differentiating (3.1) with respect to u^L, we get

$$\frac{\partial u}{\partial u^L} + \frac{\partial v}{\partial u^L} = \mu f'(u^L)$$

and

$$(\delta + \Delta t)\frac{\partial v}{\partial u^L} - \Delta t A'(u)\frac{\partial u}{\partial u^L} = 0;$$

hence

$$\frac{\partial v}{\partial u^L} = \frac{\Delta t A'(u)}{\delta + \Delta t + \Delta t A'(u)} \mu f'(u^L) \equiv \bar{\omega} \mu f'(u^L), \tag{3.8}$$

where $\bar{\omega} \in [0, 1)$ and

$$\frac{\partial u}{\partial u^L} = \mu f'(u^L)(1 - \bar{\omega}) \tag{3.9}$$

Now (3.8) and (3.9) give

$$\frac{\partial u}{\partial u^L}, \frac{\partial v}{\partial u^L} \geq 0. \tag{3.10}$$

By using the monotonicity properties (3.4), (3.7), and (3.10) of $u = u(u^L, \bar{u}, \bar{v})$ and $v = v(u^L, \bar{u}, \bar{v})$, we have

$$u(0, 0, 0) \leq u(u^L, \bar{u}, \bar{v}) \leq u(1, 1, 1)$$

and

$$v(0, 0, 0) \leq v(u^L, \bar{u}, \bar{v}) \leq v(1, 1, 1).$$

It is easily seen that $u(0, 0, 0) = v(0, 0, 0) = 0$ and $u(1, 1, 1) = v(1, 1, 1) = 1$, and then the maximum principle I of the lemma follows by induction.

3.2. Proof of II: The total-variation estimate. Recall that the total variation of $\{u_j^n\}$ at time $t = n\Delta t$ is defined by

$$\text{TV}(u^n) = \sum_j |u_{j+1}^n - u_j^n|.$$

In order to prove the TV bound, it is convenient to introduce the following notation:

$$U_j^n = u_{j+1}^n - u_j^n$$

and

$$V_j^n = v_{j+1}^n - v_j^n.$$

Then by rewriting the scheme in (2.6) in the form

$$\begin{aligned} u_j^{n+1} &= u_j^n - \mu(f(u_j^n) - f(u_{j-1}^n)) - \frac{\Delta t}{\delta}(A(u_j^{n+1}) - v_j^{n+1}), \\ v_j^{n+1} &= v_j^n + \frac{\Delta t}{\delta}(A(u_j^{n+1}) - v_j^{n+1}), \end{aligned} \tag{3.11}$$

we get

(3.12)

$$U_j^{n+1} = U_j^n - \mu f'(\tilde{u}_{j+1/2}^n) U_j^n + \mu f'(\tilde{u}_{j-1/2}^n) U_{j-1}^n - \frac{\Delta t}{\delta} A'(\hat{u}_{j+1/2}^{n+1}) U_j^{n+1} + \frac{\Delta t}{\delta} V_j^{n+1},$$

$$V_j^{n+1} = V_j^n + \frac{\Delta t}{\delta} A'(\hat{u}_{j+1/2}^{n+1}) U_j^{n+1} - \frac{\Delta t}{\delta} V_j^{n+1}. \tag{3.13}$$

Here $\tilde{u}^n_{j+1/2}$ satisfies

$$f(u^n_{j+1}) - f(u^n_j) = f'(\tilde{u}^n_{j+1/2})U^n_j \tag{3.14}$$

and $\hat{u}^n_{j+1/2}$ satisfies

$$A(u^n_{j+1}) - A(u^n_j) = A'(\hat{u}^n_{j+1/2})U^n_j. \tag{3.15}$$

We multiply (3.12) by $\sigma(U^{n+1}_j)$ and (3.13) by $\sigma(V^{n+1}_j)$ (recall that σ denotes the sign function), and get

$$\begin{aligned}
&|U^{n+1}_j| \\
&\quad\le (1 - \mu f'(\tilde{u}^n_{j+1/2})|U^n_j| + \mu f'(\tilde{u}^n_{j-1/2})|U^n_{j-1}| - \frac{\Delta t}{\delta} A'(\hat{u}^{n+1}_{j+1/2})|U^{n+1}_j| + \frac{\Delta t}{\delta}|V^{n+1}_j|,\\
&|V^{n+1}_j| \le |V^n_j| + \frac{\Delta t}{\delta} A'(\hat{u}^{n+1}_{j+1/2})|U^{n+1}_j| - \frac{\Delta t}{\delta}|V^{n+1}_j|.
\end{aligned}$$

By adding these inequalities, we get

$$\sum_j |U^{n+1}_j| + \sum_j |V^{n+1}_j| \le \sum_j |U^n_j| + \sum_j |V^n_j|,$$

and thus

$$\mathrm{TV}(u^n) + \mathrm{TV}(v^n) \le \mathrm{TV}(u^0) + \mathrm{TV}(v^0)$$

by induction.

3.3. Proof of III: Deviation from equilibrium. Note that $p = A(u) - v$ measures the deviation from equilibrium in the nonequilibrium model. Part III of Lemma 3.1 states that if we give initial data close to equilibrium for the nonequilibrium model, the solution will remain close to equilibrium for all time. To prove this, we begin by noting that

$$p^{n+1}_j - p^n_j = A'(\hat{u}^{n+1}_j)(u^{n+1}_j - u^n_j) - (v^{n+1}_j - v^n_j),$$

where $\hat{u}^{n+1}_j$ satisfies

$$A(u^{n+1}_j) - A(u^n_j) = A'(\hat{u}^{n+1}_j)(u^{n+1}_j - u^n_j).$$

Then, using the scheme in (3.11), we get

$$p^{n+1}_j = p^n_j - \mu A'(\hat{u}^{n+1}_j) f'(\tilde{u}^n_{j-1/2})(u^n_j - u^n_{j-1}) - \frac{\Delta t}{\delta}(1 + A'(\hat{u}^{n+1}_j))p^{n+1}_j, \tag{3.16}$$

where $\tilde{u}^n_{j-1/2}$ satisfies (3.14). Multiplying this equation by $\sigma(p^{n+1}_j)$, we obtain

$$|p^{n+1}_j| \le |p^n_j| + \mu M |u^n_j - u^n_{j-1}| - \frac{\Delta t}{\delta}|p^{n+1}_j|,$$

where M is a finite constant independent of Δx, Δt, and δ and where we have used the fact that

$$A'(u) \ge 0$$

for all $u \in [0,1]$; cf. (2.3). Since the total variation of u^n is bounded, we have

$$\|p^{n+1}\|_1 \leq \|p^n\|_1 + \tilde{M}\Delta t - \frac{\Delta t}{\delta}\|p^{n+1}\|_1, \tag{3.17}$$

where again $\tilde{M}$ is another finite constant independent of Δx, Δt, and δ. From (3.17), we get

$$\|p^n\|_1 \leq \tilde{M}\delta \tag{3.18}$$

provided that the estimate holds for $n = 0$. This concludes the proof of part III of the lemma.

Remark. As mentioned above, our interest in this paper is solutions which are close to equilibrium. For this purpose, estimate (3.18) is exactly what we need. It should be mentioned, however, that some further insight could be gained from (3.17). In fact, by assuming that Δt is sufficiently small, i.e., $\Delta t \leq$ constant δ, inequality (3.17) implies that

$$\|p^n\|_1 \leq M_1\delta + e^{-M_2 t_n/\delta}\|p^0\|_1, \tag{3.19}$$

where M_1 and M_2 are finite constants independent of δ, Δt, and Δx. Hence, for any initial data of bounded variation lying in the state spaces, the nonequilibrium discrete solution approaches a state close to equilibrium at an exponential rate.

3.4. Proof of IV: L^1-continuity in time. By using the property of p_j^n derived above, the L^1-continuity in time is easily verified. Consider the scheme in (3.14) and (3.15) using the definition of p_j^n:

$$\begin{aligned} u_j^{n+1} &= u_j^n - \mu(f(u_j^n) - f(u_{j-1}^n)) - \frac{\Delta t}{\delta}p_j^{n+1}, \\ v_j^{n+1} &= v_j^n + \frac{\Delta t}{\delta}p_j^{n+1}. \end{aligned} \tag{3.20}$$

Now

$$\|u^{n+1} - u^n\|_1 + \|v^{n+1} - v^n\|_1 \leq M\Delta t \sum_j |u_j^n - u_{j-1}^n| + \frac{\Delta t}{\delta}\|p^{n+1}\|_1,$$

where M is a finite constant independent of Δx, Δt, and δ. Thus, using parts I, II, and III of Lemma 3.1, we get

$$\|u^{n+1} - u^n\|_1 + \|v^{n+1} - v^n\|_1 = O(\Delta t),$$

and then IV follows using the triangle inequality.

3.5. Proof of V: Stability in L^1. The proof of the L^1-stability is quite similar to the proof of the TV bound and is therefore omitted.

3.6. The discrete entropy inequalities. Finally, we would like to show that the solutions of the difference scheme in (2.6) satisfy an entropy inequality.

LEMMA 3.2. *Suppose that (u_j^0, v_j^0) satisfies (2.9), and let $\phi, \psi \in \mathcal{D}_+(T)$ for some $T > 0$. Furthermore, let N be a positive integer such that $t_N \leq T$, and let E :*

$[0,1] \to \mathcal{R}$ be a convex C^∞ entropy function with an associated entropy flux F, i.e., $F : [0,1] \to \mathcal{R}$ satisfies $F' = E'f'$. Then the following inequalities hold:

$$\begin{aligned} &\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathcal{Z}} \left(\frac{\phi(x_j,t_{n+1}) - \phi(x_j,t_n)}{\Delta t} E(u_j^{n+1}) + \frac{\phi(x_{j+1},t_n) - \phi(x_j,t_n)}{\Delta x} F(u_j^n) \right) \\ &\qquad + \Delta x \sum_{j\in\mathcal{Z}} \left(E(u_j^0)\phi(x_j,0) - E(u_j^N)\phi(x_j,t_N) \right) \qquad (3.21) \\ &\qquad \geq \left(\frac{\Delta t}{\delta}\right) \sum_{n=0}^{N-1} (\Delta x) \sum_{j\in\mathcal{Z}} E'(u_j^{n+1})\phi(x_j,t_n)(A(u_j^{n+1}) - v_j^{n+1}) \end{aligned}$$

and

$$\begin{aligned} &\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathcal{Z}} \left(\frac{\psi(x_j,t_{n+1}) - \psi(x_j,t_n)}{\Delta t} E(v_j^{n+1}) \right) \\ &\qquad + \Delta x \sum_{j\in\mathcal{Z}} \left(E(v_j^0)\psi(x_j,0) - E(v_j^N)\psi(x_j,t_N) \right) \qquad (3.22) \\ &\qquad \geq -\left(\frac{\Delta t}{\delta}\right) \sum_{n=0}^{N-1} (\Delta x) \sum_{j\in\mathcal{Z}} E'(v_j^{n+1})\psi(x_j,t_n)(A(u_j^{n+1}) - v_j^{n+1}). \end{aligned}$$

Proof. It will be sufficient to establish inequality (3.21) since (3.22) will follow by completely similar arguments. From the difference scheme in (2.6), we observe that

$$\begin{aligned} u_j^{n+1} - k = u_j^n - k - \mu(f(u_j^n) - f(k)) + \mu(f(u_{j-1}^n) - f(k)) \\ - \left(\frac{\Delta t}{\delta}\right)(A(u_j^{n+1}) - v_j^{n+1}) \end{aligned} \qquad (3.23)$$

for any $k \in [0,1]$. By multiplying this equality by $\sigma(u_j^{n+1} - k)$ and observing that the CFL condition (2.8) implies that

$$|u_j^n - k - \mu(f(u_j^n) - f(k))| = |u_j^n - k| - \mu|f(u_j^n) - f(k)|,$$

we obtain

$$\begin{aligned} |u_j^{n+1} - k| \leq |u_j^n - k| - \mu|f(u_j^n) - f(k)| + \mu|f(u_{j-1}^n) - f(k)| \\ - \left(\frac{\Delta t}{\delta}\right) \sigma(u_j^{n+1} - k)(A(u_j^{n+1}) - v_j^{n+1}). \end{aligned}$$

If we multiply this final inequality by $\phi(x_j,t_n)$, sum over $0 \leq n \leq N-1$ and $j \in \mathcal{Z}$, and apply summation by parts with respect to time and space, the following inequality appears:

$$\begin{aligned} &\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathcal{Z}} \left(\frac{\phi(x_j,t_{n+1}) - \phi(x_j,t_n)}{\Delta t} |u_j^{n+1} - k| \right. \\ &\qquad \left. + \frac{\phi(x_{j+1},t_n) - \phi(x_j,t_n)}{\Delta x} |f(u_j^n) - f(k)| \right) \qquad (3.24) \\ &\qquad + \Delta x \sum_{j\in\mathcal{Z}} \left(|u_j^0 - k|\phi(x_j,0) - |u_j^N - k|\phi(x_j,t_N) \right) \qquad (3.25) \end{aligned}$$

$$\geq \left(\frac{\Delta t}{\delta}\right) \sum_{n=0}^{N-1} (\Delta x) \sum_{j \in \mathcal{Z}} \sigma(u_j^{n+1} - k)\phi(x_j, t_n)(A(u_j^{n+1}) - v_j^{n+1}).$$

Now let $E_m : I \to \mathcal{R}$ be a convex piecewise-linear function of the form

$$E_m(u) = \beta_0(u - k_0) + \sum_{i=1}^{m} \beta_i |u - k_i|, \tag{3.26}$$

where $\beta_i \geq 0$ for $i = 1, 2, \ldots, m$, and let F_m be the corresponding flux given by

$$F_m(u) = \beta_0(f(u) - f(k_0)) + \sum_{i=1}^{m} \beta_i |f(u) - f(k_i)|.$$

Observe that $F_m'(u) = E_m'(u)f'(u)$ and that equality (3.23) and inequality (3.25) imply that the desired inequality (3.21) holds if the smooth functions E and F are replaced by the polygonal functions E_m and F_m. However, for any smooth convex function E, we can choose E_m of the form of (3.26) such that E_m and E_m' converge uniformly on $[0, 1]$ as $m \to \infty$ to E and E', respectively, and hence we obtain (3.21). □

4. Properties of the entropy solutions.

4. Properties of the entropy solutions. In this section, we shall conclude the proof of Theorem 2.1. The properties of the entropy solutions of system (2.2) will be derived from the corresponding properties derived for the finite-difference scheme in (2.6) above. First, we will establish the existence of entropy solutions for (2.2) by a limit argument. Thereafter, uniqueness and continuous dependence with respect to the initial data in L^1 are proved. From a proper application of Helly's theorem, the following lemma can be derived by standard arguments from parts I–IV of Lemma 3.1 (cf. [16] or [23, Chapter 16]).

LEMMA 4.1. *Suppose that (u^0, v^0) satisfies the requirements in (2.4) above. Then as the mesh parameters Δx and Δt tend to zero, there is a subsequence $\{(u_\Delta, v_\Delta)\}$ of the family of approximate solutions generated by (2.6) that converges in $(L^1_{\text{loc}}(\mathcal{R} \times \mathcal{R}_0^+))^2$ to a pair of functions (u, v). Furthermore, $u(\cdot, t), v(\cdot, t) \in \text{BV}$ for all $t \geq 0$, $(u(x, t), v(x, t)) \in \mathcal{S}$ for $(x, t) \in \mathcal{R} \times \mathcal{R}_0^+$, and the estimates*

$$\text{TV}(u(\cdot, t)) + \text{TV}(v(\cdot, t)) \leq \text{TV}(u^0) + \text{TV}(v^0), \tag{4.1}$$

$$\|p(\cdot, t)\|_1 \leq M\delta, \tag{4.2}$$

$$\|u(\cdot, t) - u(\cdot, \tau)\|_1 + \|v(\cdot, t) - v(\cdot, \tau)\|_1 \leq M|t - \tau| \tag{4.3}$$

hold, where $p(\cdot, t) = A(u(\cdot, t)) - v(\cdot, t)$ and M is independent of t and δ. □

Since the discrete solutions satisfy the entropy inequalities (3.21) and (3.22), we also easily derive that the limit (u, v) is an entropy solution of (2.2).

LEMMA 4.2. *Assume that (u^0, v^0) satisfies the requirements in (2.4), and let (u, v) be the pair of functions constructed in Lemma 4.1. Then (u, v) is an entropy solution of (2.2).*

Proof. We have to show that (u, v) satisfy the variational inequality (2.5) for any $\varphi, \psi \in \mathcal{D}_+(T)$ and $(k, q) \in \mathcal{S}$. However, by letting $\Delta t, \Delta x \to 0$ in (3.21), it follows that for any smooth entropy/entropy flux pair (E, F),

$$\int_0^T \int_{\mathcal{R}} [E(u)\varphi_t + F(u)\varphi_x] dx\, dt$$

$$\begin{aligned} &+\int_{\mathcal{R}}[E(u^0)\varphi(x,0)-E(u(x,T))\varphi(x,T)]dx \\ &\geq \frac{1}{\delta}\int_0^T\int_{\mathcal{R}}E'(u)\varphi(A(u)-v)dx\,dt. \end{aligned} \tag{4.4}$$

Hence, by choosing a sequence of smooth entropy/entropy flux pairs (E_θ, F_θ) such that as $\theta \to 0$,

$$E_\theta(u) \to |u-k| \quad \text{and} \quad E'_\theta \to \sigma(u-k)$$

pointwise, the dominated-convergence theorem implies that

$$\begin{aligned} &\int_0^T\int_{\mathcal{R}}[|u-k|\varphi_t + |f(u)-f(k)|\varphi_x]dx\,dt \\ &\qquad +\int_{\mathcal{R}}[|u^0-k|\varphi(x,0)-|u(x,T)-k|\varphi(x,T)]dx \\ &\qquad \geq \frac{1}{\delta}\int_0^T\int_{\mathcal{R}}\sigma(u-k)\varphi(A(u)-v)dx\,dt \end{aligned}$$

for all $k \in [0,1]$ and $\varphi \in \mathcal{D}_+(T)$. The rest of the desired inequality (2.5) can be derived from (3.22) by similar arguments. □

In order to show that the entropy solutions are unique, we will use arguments inspired by Kruzkov [9], Kuznetsov [10], and Lucier [14]. For any $\theta \in (0,1]$, we introduce the mollifier function ω_θ on $\mathcal{R}$ given by

$$\omega_\theta(x) = \frac{1}{\theta}\Omega\left(\frac{x}{\theta}\right),$$

where $\Omega : \mathcal{R} \to \mathcal{R}$ is a nonnegative, symmetric C^∞-function with support in $[-1,1]$ and satisfying

$$\int_{\mathcal{R}}\Omega(x)dx = 1.$$

Hence

$$\int_{\mathcal{R}}\omega_\theta(x)dx = 1$$

and $\operatorname{supp}(\omega_\theta) \subset [-\theta,\theta]$. Furthermore, we define a smooth approximation σ_θ of the sign function σ by

$$\sigma_\theta(x) = -1 + 2\int_{-\infty}^x \omega_\theta(y)dy$$

and a smooth approximation μ_θ of the absolute-value function by

$$\mu_\theta(x) = \theta + \int_{-\theta}^x \sigma_\theta(y)dy.$$

We observe that ${\mu_\theta}' = \sigma_\theta$ and ${\sigma_\theta}' = 2\omega_\theta$. Also,

$$\begin{aligned} \sigma_\theta(x) &= \sigma(x) \quad \text{for } |x| \geq \theta, \\ \mu_\theta(x) &= |x| \quad \text{for } |x| \geq \theta, \end{aligned}$$

and

$$|\mu_\theta(x) - |x|| \le \theta \quad \text{for all } x \in \mathcal{R}.$$

The next result shows that the entropy solutions of (2.2) are unique and depend continuously, independently of δ, on the initial data in L^1.

LEMMA 4.3. *Let (u, v) and $(\bar{u}, \bar{v})$ be two entropy solutions of* (2.2) *with initial data (u^0, v^0) and $(\bar{u}^0, \bar{v}^0)$, respectively, satisfying the requirements in* (2.4). *Then*

$$\|u(\cdot, t) - \bar{u}(\cdot, t)\|_1 + \|v(\cdot, t) - \bar{v}(\cdot, t)\|_1 \le \|u^0 - \bar{u}^0\|_1 + \|v^0 - \bar{v}^0\|_1 \quad \text{for all } t \ge 0.$$

Proof. The result follows by generalizing Kuznetsov's argument in [10] for scalar conservation laws. Let $T > 0$ be given and choose $(k, q) = (\bar{u}(y, \tau), \bar{v}(y, \tau))$ and $\varphi(x, t) = \psi(x, t) = \omega_\theta(x - y)\omega_\theta(t - \tau)$ in (2.5) for the solution (u, v). Integrating the result over $\mathcal{R} \times [0, T]$ with respect to y and τ, we obtain

(4.5)
$$\begin{aligned}
&\int_0^T \int_{\mathcal{R}} \int_0^T \int_{\mathcal{R}} (|u - \bar{u}| + |v - \bar{v}|)\omega_\theta{}'(t - \tau)\omega_\theta(x - y) dx\, dt\, dy\, d\tau \\
&\quad + \int_0^T \int_{\mathcal{R}} \int_0^T \int_{\mathcal{R}} |f(u) - f(\bar{u})| \omega_\theta(t - \tau)\omega_\theta{}'(x - y) dx\, dt\, dy\, d\tau \\
&\quad + \int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} [|u^0(x) - \bar{u}| + |v^0(x, t) - \bar{v}|]\omega_\theta(\tau)\omega_\theta(x - y) dx\, dy\, d\tau \\
&\quad - \int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} [|u(x, T) - \bar{u}| + |v(x, T) - \bar{v}|]\omega_\theta(T - \tau)\omega_\theta(x - y) dx\, dy\, d\tau \\
&\ge \frac{1}{\delta} \int_0^T \int_{\mathcal{R}} \int_0^T \int_{\mathcal{R}} (\sigma(u - \bar{u}) - \sigma(v - \bar{v}))(A(u) - v)\omega_\theta(x - y)\omega_\theta(t - \tau) dx\, dt\, dy\, d\tau,
\end{aligned}$$

where $u = u(x, t)$, $v = v(x, t)$, $\bar{u} = \bar{u}(y, \tau)$, and $\bar{v} = \bar{v}(y, \tau)$. By performing a similar operation on inequality (2.5) for the solution $(\bar{u}, \bar{v})$, but where we reverse the role of the variables (x, t) and (y, τ), and by adding the inequality obtained to (4.5), we observe that the terms which contain derivatives will cancel out. Hence we obtain an inequality of the form

(4.6)
$$\begin{aligned}
&R(\theta) - L(\theta) \\
&\quad \ge \frac{1}{\delta} \int_0^T \int_{\mathcal{R}} \int_0^T \int_{\mathcal{R}} (\sigma(u - \bar{u}) - \sigma(v - \bar{v}))(A(u) - A(\bar{u}) \\
&\qquad - (v - \bar{v}))\omega_\theta(x - y)\omega_\theta(t - \tau) dx\, dt\, dy\, d\tau,
\end{aligned}$$

where

$$\begin{aligned}
R(\theta) = &\int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} (|u^0(x) - \bar{u}| + |v^0(x) - \bar{v}|)\omega_\theta(x - y)\omega_\theta(\tau) dx\, dy\, d\tau \\
&+ \int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} (|u - \bar{u}^0(y)| + |v - \bar{v}^0(y)|)\omega_\theta(x - y)\omega_\theta(t) dx\, dy\, dt
\end{aligned}$$

and

$$\begin{aligned}
L(\theta) = &\int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} (|u(x, T) - \bar{u}| + |v(x, T) - \bar{v}|)\omega_\theta(T - \tau)\omega_\theta(x - y) dx\, dy\, d\tau \\
&+ \int_0^T \int_{\mathcal{R}} \int_{\mathcal{R}} (|u - \bar{u}(y, T)| + |v - \bar{v}(y, T)|)\omega_\theta(T - t)\omega_\theta(x - y) dx\, dy\, dt.
\end{aligned}$$

Due to the monotonicity of $A = A(u)$ (cf. (2.3)), the integrand on the right-hand side of (4.6) is nonnegative, and thus we get

$$L(\theta) \leq R(\theta). \tag{4.7}$$

By using the properties in (4.1) and (4.3) for both pairs of solutions, it follows easily (cf., e.g., [26]) that

$$|R(\theta) - (\|u^0 - \bar{u}^0\|_1 + \|v^0 - \bar{v}^0\|_1)| \leq M\theta$$

and

$$|L(\theta) - (\|u(\cdot, T) - \bar{u}(\cdot, T)\|_1 + \|v(\cdot, T) - \bar{v}(\cdot, T)\|_1)| \leq M\theta,$$

where M is independent of θ. Thus the desired result follows from (4.7) by letting $\theta \to 0$. □

Remarks. (1) Note that Lemmas 4.1, 4.2, and 4.3 prove Theorem 2.1. In particular, the uniqueness result of Lemma 4.3 implies that the complete sequence $\{(u_\Delta, v_\Delta)\}$ converges towards (u, v).

(2) In the proof of Theorem 2.1, we have not used the assumption on A'' given in (2.3).

(3) The stability result of Lemma 4.3 could also be proved by appealing to the similar property of the discrete scheme; cf. part V of Lemma 3.1. However, in order to prove both stability *and uniqueness* of the entropy solution, the argument presented above is needed.

5. A regularized model. The purpose of the final two sections of this paper is to analyze the convergence as δ tends to zero of the solutions of the model in (2.2) to the corresponding solutions of the equilibrium model given in (2.1). The main purpose of the discussion is to establish that the difference between these solutions, measured in the L^1-norm, is bounded by $O(\delta^{1/3})$.

In order to study this convergence, a regularized model is introduced. For any $\epsilon, \delta \in (0, 1]$, consider the pure initial-value problem

$$\begin{aligned} u_t^\epsilon + f(u^\epsilon)_x &= -\frac{1}{\delta}(A(u^\epsilon) - v^\epsilon) + \epsilon u_{xx}^\epsilon, \\ v_t^\epsilon &= \frac{1}{\delta}(A(u^\epsilon) - v^\epsilon) + \epsilon v_{xx}^\epsilon, \\ u^\epsilon(x, 0) = u^{\epsilon,0}(x), &\qquad v^\epsilon(x, 0) = v^{\epsilon,0}(x), \end{aligned} \tag{5.1}$$

where f and A are the given functions of u satisfying (2.3).

Since system (5.1) contains proper diffusion terms, the solutions will in general be smooth functions of x and t. In order to state a suitable regularity result for the system, we let $H^m(\mathcal{R})$, $m \geq 0$, denote the L^2-based Sobolev spaces of order m on $\mathcal{R}$ and we let $H^\infty(\mathcal{R}) = \bigcap_{m \geq 0} H^m(\mathcal{R})$.

Hence $H^\infty(\mathcal{R})$ consists of all C^∞-functions with the property that any derivative is in $L^2(\mathcal{R})$. In particular, if $u \in H^\infty(\mathcal{R})$, then

$$\frac{\partial^j}{\partial x^j} u(x) \to 0 \quad \text{as } x \to \pm\infty \tag{5.2}$$

for any $j \geq 0$.

In order to avoid technical difficulties with respect to smoothness, we will simply assume that the initial functions $u^{\epsilon,0}$ and $v^{\epsilon,0}$ of (5.1) are in $H^\infty(\mathcal{R})$. The following regularity result can then be established.

Lemma 5.1. *Assume that* $(u^{\epsilon,0}(x), v^{\epsilon,0}(x)) \in \mathcal{S}$ *for all* $x \in \mathcal{R}$ *and that* $u^{\epsilon,0}, v^{\epsilon,0} \in H^\infty(\mathcal{R})$. *Then there exists a unique classical solution* (u^ϵ, v^ϵ) *of* (5.1) *such that*

$$\frac{\partial^j u^\epsilon}{\partial t^j}(\cdot, t),\ \frac{\partial^j v^\epsilon}{\partial t^j}(\cdot, t) \in H^\infty(\mathcal{R})$$

for all $t \geq 0$ *and integers* $j \geq 0$. *Furthermore,*

$$(u^\epsilon(x,t), v^\epsilon(x,t)) \in \mathcal{S}$$

for all $x \in \mathcal{R}$ *and* $t \geq 0$. □

The regularity part of this lemma is rather standard and a proof can be found, for example, in [8] (cf. Chapter 5 of [8]), while the invariant region part follows from the assumptions on A and the result of Chueh, Conley, and Smoller [4], (cf. Theorem 4.4. of [4]).

In section 3, independently of the relaxation parameter δ and the discretization parameters Δx and Δt, we derived bounds for the total variation of the solution of the discrete scheme (2.6). We will need similar bounds for the solution of the regularized model (5.1), i.e., TV bounds independent of the parameters ϵ and δ. We observe that if $u \in H^\infty(\mathcal{R})$, then $u \in \mathrm{BV}$ if and only if $u_x \in L^1(\mathcal{R})$. Note that for these functions, $\mathrm{TV}(u) = \|u_x\|_1$.

Lemma 5.2. *Assume that* $(u^{\epsilon,0}(x), v^{\epsilon,0}(x)) \in \mathcal{S}$ *for all* $x \in \mathcal{R}$ *and that* $u^{\epsilon,0}, v^{\epsilon,0} \in H^\infty(\mathcal{R}) \bigcap \mathrm{BV}$. *If* (u^ϵ, v^ϵ) *is the solution of* (5.1), *then* $u^\epsilon(\cdot,t), v^\epsilon(\cdot,t) \in H^\infty(\mathcal{R}) \bigcap \mathrm{BV}$ *for all* $t \geq 0$ *and*

$$\|u_x^\epsilon(\cdot,t)\|_1 + \|v_x^\epsilon(\cdot,t)\|_1 \leq \|u_x^{\epsilon,0}\|_1 + \|v_x^{\epsilon,0}\|_1. \tag{5.3}$$

Furthermore, if the initial data is chosen such that $u_t^\epsilon(\cdot,0), v_t^\epsilon(\cdot,0) \in L^1(\mathcal{R})$, *then* $u_t^\epsilon(\cdot,t), v_t^\epsilon(\cdot,t) \in L^1(\mathcal{R})$ *for all* $t \geq 0$ *and*

$$\|u_t^\epsilon(\cdot,t)\|_1 + \|v_t^\epsilon(\cdot,t)\|_1 \leq \|u_t^\epsilon(\cdot,0)\|_1 + \|v_t^\epsilon(\cdot,0)\|_1. \tag{5.4}$$

Proof. For notational convenience, we will throughout this proof denote the solution of (5.1) by (u, v), i.e., the explicit dependence on the regularization parameter ϵ is suppressed. Let $\mu_\theta : \mathcal{R} \to \mathcal{R}$, $\theta \in (0,1]$, be the smooth approximation of the absolute-value function introduced in section 4. Hence ${\mu_\theta}' = \sigma_\theta$, where σ_θ is the smooth approximation of the sign function, while

$${\mu_\theta}'' = {\sigma_\theta}' = 2\omega_\theta.$$

Now consider system (5.1) and differentiate both equations with respect to x. We then obtain the following system:

$$\begin{aligned}(u_x)_t + (f'(u)u_x)_x &= -\frac{1}{\delta}(A'(u)u_x - v_x) + \epsilon(u_x)_{xx},\\ (v_x)_t &= \frac{1}{\delta}(A'(u)u_x - v_x) + \epsilon(v_x)_{xx}.\end{aligned} \tag{5.5}$$

Furthermore, multiply the first equation of (5.5) by $\sigma_\theta(u_x)$ and the second equation by $\sigma_\theta(v_x)$ and integrate over $(-M, M) \times (0, T)$ for $M, T > 0$. If we add the results from the two equations, we obtain

$$\tag{5.6}\begin{aligned}&\int_{-M}^{M}(\mu_\theta(u_x(x,T))+\mu_\theta(v_x(x,T)))dx+\int_0^T\int_{-M}^{M}(f'(u)u_x)_x\sigma_\theta(u_x)dx\ dt\\&=\int_{-M}^{M}(\mu_\theta(u_x^0(x))+\mu_\theta(v_x^0(x)))dx-\frac{1}{\delta}\int_0^T\int_{-M}^{M}(A'(u)u_x-v_x)(\sigma_\theta(u_x)-\sigma_\theta(v_x))dx\ dt\\&\qquad+\epsilon\int_0^T\int_{-M}^{M}(u_{xxx}\sigma_\theta(u_x)+v_{xxx}\sigma_\theta(v_x))dx\ dt.\end{aligned}$$

We observe that since $|\sigma_\theta(r)|\le 1$ and since $\sigma_\theta(r)\to\sigma(r)$ for all $r\in\mathcal{R}$, it follows from the dominated-convergence theorem that

$$\tag{5.7}\begin{aligned}\lim_{\theta\to 0}-\int_0^T\int_{-M}^{M}&(A'(u)u_x-v_x)(\sigma_\theta(u_x)-\sigma_\theta(v_x))dx\ dt\\&=-\int_0^T\int_{-M}^{M}(A'(u)u_x-v_x)(\sigma(u_x)-\sigma(v_x))dx\ dt\le 0.\end{aligned}$$

By using integration by parts with respect to x, we derive

$$\begin{aligned}&\int_0^T\int_{-M}^{M}(f'(u)u_x)_x\sigma_\theta(u_x)dx\ dt\\&\quad=\int_0^T[f'(u)u_x\sigma_\theta(u_x)]\mid_{x=-M}^{M}dt-2\int_0^T\int_{-M}^{M}f'(u)u_xu_{xx}\omega_\theta(u_x)dx\ dt.\end{aligned}$$

Since $r\omega_\theta(r)$ is uniformly bounded and since for any $r\in\mathcal{R}$, $\lim_{\theta\to 0}r\omega_\theta(r)=0$, it follows again from the dominated-convergence theorem that

$$\lim_{\theta\to 0}\int_0^T\int_{-M}^{M}f'(u)u_xu_{xx}\omega_\theta(u_x)dx\ dt=0.$$

Hence (5.2) implies that

$$\tag{5.8}\lim_{M\to\infty}\lim_{\theta\to 0}\int_0^T\int_{-M}^{M}(f'(u)u_x)_x\sigma_\theta(u_x)dx\ dt=0.$$

Finally, we observe that integration by parts with respect to x implies that

$$\int_0^T\int_{-M}^{M}(u_x)_{xx}\sigma_\theta(u_x)dx\ dt=\int_0^T u_{xx}\sigma_\theta(u_x)\mid_{x=-M}^{M}dt-2\int_0^T\int_{-M}^{M}u_{xx}^2\omega_\theta(u_x)dx\ dt.$$

Since $\omega_\theta(r)\ge 0$ and by treating the v term in the same way, we therefore obtain

$$\lim_{M\to\infty}\limsup_{\theta\to 0}\epsilon\int_0^T\int_{-M}^{M}((u_x)_{xx}\sigma_\theta(u_x)+(v_x)_{xx}\sigma_\theta(v_x))dx\ dt\le 0.$$

Together with (5.6), (5.7), and (5.8), this implies that

$$\lim_{M\to\infty}\int_{-M}^{M}(|u_x(x,T)|+|v_x(x,T)|)dx\le\|u_x^0\|_1+\|v_x^0\|_1,$$

and hence $u(\cdot,T), v(\cdot,T) \in \mathrm{BV}$ and (5.3) holds. Inequality (5.4) can be derived by a completely analogous argument by differentiating system (5.1) with respect to t instead of x. The details of this derivation are therefore omitted. □

As above, we need to measure the deviation from equilibrium,

$$p^\epsilon = A(u^\epsilon) - v^\epsilon.$$

LEMMA 5.3. *Assume that the initial data $u^{\epsilon,0}$ and $v^{\epsilon,0}$ of* (5.1) *satisfy all of the assumptions given in Lemma* 5.2. *Furthermore, assume that $\|p^\epsilon(\cdot,0)\|_1 \le M_0\delta$ and $\|p^\epsilon_x(\cdot,0)\|_1 \le M_0\delta/\epsilon$, where M_0 is independent of ϵ and δ. Then there is a constant M, independent of ϵ and δ, such that*

$$\|p^\epsilon(\cdot,t)\|_1 \le M\delta \tag{5.9}$$

and

$$\|p^\epsilon_x(\cdot,t)\|_1 \le \frac{M\delta}{\epsilon} \tag{5.10}$$

for all $t \ge 0$.

Proof. As above, we suppress the explicit dependence of ϵ throughout the proof. From the definition of the function p and from (5.1), it follows that

$$p_t = A'(u)u_t - v_t = A'(u)\left(\epsilon u_{xx} - \frac{1}{\delta}p - f(u)_x\right) - \epsilon v_{xx} - \frac{1}{\delta}p.$$

Since

$$p_{xx} = A'(u)u_{xx} + A''(u){u_x}^2 - v_{xx},$$

the expression above can be written as follows:

$$p_t = \epsilon p_{xx} - \frac{1}{\delta}(1 + A'(u))p - \epsilon A''(u){u_x}^2 - A'(u)f(u)_x. \tag{5.11}$$

Observe that if we multiply the first equation of (5.1) by $u(x,t)$ and integrate with respect to x, we obtain

$$\epsilon\int_{\mathcal{R}} {u_x}^2(x,t)dx \le \frac{1}{\delta}\int_{\mathcal{R}} |p(x,t)||u(x,t)|dx + \int_{\mathcal{R}} |u_t(x,t)||u(x,t)|dx.$$

Hence the results of Lemmas 5.1 and 5.2 imply that there is a constant M, independent of ϵ, δ, and t, such that

$$\epsilon\int_{\mathcal{R}} {u_x}^2(x,t)dx \le \frac{\|p(\cdot,t)\|_1}{\delta} + M.$$

If we multiply (5.11) by $\sigma_\theta(p)$, integrate with respect to x, and let θ tend to zero, we obtain

$$\frac{d}{dt}\|p(\cdot,t)\|_1 \le M - \frac{1-\alpha}{\delta}\|p(\cdot,t)\|_1,$$

where $0 \le \alpha < 1$ (cf. (2.3)). Gronwall's lemma therefore implies that

$$\begin{aligned}
&\|p(\cdot,t)\|_1 \\
&\quad \le \exp\left(-\frac{(1-\alpha)t}{\delta}\right)\|p(\cdot,0)\|_1 + \left(\frac{M\delta}{(1-\alpha)}\right)\left(1 - \exp\left(-\frac{(1-\alpha)t}{\delta}\right)\right) \\
&\quad \le \|p(\cdot,0)\|_1 + \frac{M\delta}{1-\alpha}.
\end{aligned}$$

We have therefore established (5.9).

In order to derive (5.10), we differentiate (5.11) with respect to x and obtain the equation

$$(p_x)_t = \epsilon(p_x)_{xx} - \frac{1}{\delta}(1 + A'(u))p_x + R, \tag{5.12}$$

where $R = R(x,t)$ is given by

$$\begin{aligned} R = &-\epsilon(A'''(u){u_x}^3 + 2A''(u)u_x u_{xx}) \\ &- \frac{1}{\delta}A''(u)u_x p - A''(u)f'(u){u_x}^2 - A'(u)(f'(u)u_{xx} - f''(u){u_x}^2). \end{aligned}$$

Now observe that $\|u_t(\cdot,t)\|_1, \|u_x(\cdot,t)\|_1$, and $(1/\delta)\|p(\cdot,t)\|_1$ are bounded independently of ϵ, δ, and t. Hence the first equation of (5.1) implies that $\epsilon\|u_{xx}\|_1$ admits a similar bound, and since

$$\lim_{x\to\pm\infty} u_x(x,t) = 0,$$

we obtain

$$\|u_x(\cdot,t)\|_\infty \le \frac{M}{\epsilon}, \tag{5.13}$$

where M is independent of ϵ, δ, and t. However, from (5.13) and the earlier bounds (5.3) and (5.9), we derive that

$$\|R(\cdot,t)\|_1 \le \frac{M}{\epsilon} \tag{5.14}$$

for all $t \ge 0$. Arguing as we did above, we therefore obtain from (5.12) that

$$\frac{d}{dt}\|p_x(\cdot,t)\|_1 \le \frac{M}{\epsilon} - \frac{1}{\delta}\|p_x(\cdot,t)\|_1,$$

and hence (5.10) follows from Gronwall's lemma. □

6. The rate of convergence. The purpose of this final section of the paper is to complete the proof of Theorem 2.2. Let (u,v) be the entropy solution of (2.2) with initial data (u^0,v^0). As in Theorem 2.1, we assume that the initial data satisfy the requirements in (2.4), i.e., $(u^0(x),v^0(x)) \in \mathcal{S}$ for $x \in \mathcal{R}$, $u^0, v^0 \in \mathrm{BV}$, and

$$\lim_{x\to\pm\infty} u^0(x) = \lim_{x\to\pm\infty} v^0(x) = 0$$

and that

$$\|p^0\|_1 = \|A(u^0) - v^0\|_1 \le M\delta,$$

where M is independent of δ. Furthermore, let w be the entropy solution of the equilibrium model (2.1) with $w(x,0) = u^0(x)$.

Our goal is to show that

$$\|u(\cdot,t) - w(\cdot,t)\|_1 \le M\delta^{1/3} \quad \text{for } 0 \le t \le T, \tag{6.1}$$

where M is independent of δ.

In order to establish the desired estimate (6.1), we shall rely on properties of the regularized model (5.1). Throughout this section, we assume that $\delta \leq \epsilon$. In fact, at the end of this section, we will choose $\epsilon = \delta^{2/3}$.

Let $\{u^{\epsilon,0}\}$ be a sequence of functions in $H^\infty(\mathcal{R})$ such that

$$\begin{gathered}\|u^{\epsilon,0} - u^0\|_1 \leq M\epsilon,\\ \mathrm{TV}(u^{\epsilon,0}) \leq \mathrm{TV}(u^0),\\ \|u^{\epsilon,0}_{xx}\|_1 \leq \frac{M}{\epsilon},\end{gathered} \tag{6.2}$$

where the constant M is independent of ϵ and δ. Such a sequence can be constructed by a standard averaging procedure using the mollifier function ω_ϵ. Furthermore, let

$$v^{\epsilon,0} = A(u^{\epsilon,0}); \tag{6.3}$$

then $p^{\epsilon,0} \equiv 0$ and

$$\begin{gathered}(u^{\epsilon,0}(x), v^{\epsilon,0}(x)) \in \mathcal{S} \quad \text{for all } x \in \mathcal{R},\\ \|v^{\epsilon,0} - v^0\|_1 \leq M\epsilon,\\ \mathrm{TV}(v^{\epsilon,0}) \leq M,\\ \|v^{\epsilon,0}_{xx}\|_1 \leq \frac{M}{\epsilon}.\end{gathered} \tag{6.4}$$

Let (u^ϵ, v^ϵ) denote the solution of system (5.1) corresponding to the initial data $(u^{\epsilon,0}, v^{\epsilon,0})$. We observe that (6.2) and (6.4) imply that $\|u^\epsilon_t(\cdot,0)\|_1$ and $\|v^\epsilon_t(\cdot,0)\|_1$ are bounded independently of ϵ and δ. Hence the initial data $(u^{\epsilon,0}, v^{\epsilon,0})$ satisfy all of the assumption of the Lemmas 5.1–5.3.

LEMMA 6.1. *Let u^0, v^0, $\{u^{\epsilon,0}\}$, and $\{v^{\epsilon,0}\}$ satisfy* (2.4), (6.2), *and* (6.3). *Then for any $T > 0$, there is a constant M, independent of ϵ and δ, such that*

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\|_1 + \|v^\epsilon(\cdot,t) - v(\cdot,t)\|_1 \leq M\epsilon^{1/2} \quad \textit{for } 0 \leq t \leq T.$$

Proof. The proof is based on the characterization in (2.5) of the entropy solution of (2.2) and a corresponding variational inequality for the regularized system (5.1). We first observe that for any $(k,q) \in \mathcal{S}$, the following system holds

$$\begin{gathered}(u^\epsilon - k)_t + (f(u^\epsilon) - f(k))_x = -\frac{1}{\delta}(A(u^\epsilon) - v^\epsilon) + \epsilon u^\epsilon_{xx},\\ (v^\epsilon - q)_t = \frac{1}{\delta}(A(u^\epsilon) - v^\epsilon) + \epsilon v^\epsilon_{xx}.\end{gathered} \tag{6.5}$$

Let $\varphi, \psi \in \mathcal{D}_+(T)$. For any $\theta \in (0,1)$, multiply the first equation of (6.5) by $\sigma_\theta(u^\epsilon - k)\varphi$ and the second equation by $\sigma_\theta(v^\epsilon - k)\psi$, add the two equations, and integrate over $\mathcal{R} \times (0,T)$. We then obtain

$$\begin{aligned}\int_0^T \int_{\mathcal{R}} &[\mu_\theta(u^\epsilon - k)\varphi_t + F_\theta(u^\epsilon, k)\varphi_x + \mu_\theta(v^\epsilon - q)\psi_t] dx\, dt\\ &+ \int_{\mathcal{R}} [\mu_\theta(u^{\epsilon,0}(x) - k)\varphi(x,0) + \mu_\theta(v^{\epsilon,0}(x) - q)\psi(x,0)] dx\\ &- \int_{\mathcal{R}} [\mu_\theta(u^\epsilon(x,T) - k)\varphi(x,T) + \mu_\theta(v^\epsilon(x,T) - q)\psi(x,T)] dx\end{aligned} \tag{6.6}$$

$$= \frac{1}{\delta}\int_0^T\int_{\mathcal{R}} (\sigma_\theta(u^\epsilon - k)\varphi - \sigma_\theta(v^\epsilon - q)\psi)(A(u^\epsilon) - v^\epsilon)dx\ dt$$
$$- \epsilon\int_0^T\int_{\mathcal{R}} [\sigma_\theta(u^\epsilon - k)\varphi u^\epsilon_{xx} + \sigma_\theta(v^\epsilon - q)\psi v^\epsilon_{xx}]dx\ dt,$$

where the flux function $F_\theta(u, k)$ satisfies

$$\frac{dF_\theta(u,k)}{du} = \sigma_\theta(u-k)f'(u), \quad F_\theta(k,k) = 0.$$

We observe that

$$\begin{aligned} -\epsilon&\int_0^T\int_{\mathcal{R}} \sigma_\theta(u^\epsilon - k)u^\epsilon_{xx}\varphi dx\ dt \\ &= -\epsilon\int_0^T\int_{\mathcal{R}} (\sigma_\theta(u^\epsilon - k)u^\epsilon_x)_x\varphi dx\ dt + 2\epsilon\int_0^T\int_{\mathcal{R}} \omega_\theta(u^\epsilon - k)(u^\epsilon_x)^2\varphi dx\ dt \\ &\geq \epsilon\int_0^T\int_{\mathcal{R}} \sigma_\theta(u^\epsilon - k)u^\epsilon_x\varphi_x dx\ dt \end{aligned}$$

and that a similar inequality can be derived for the corresponding v term. By letting $\theta \to 0$ in (6.6), we therefore obtain that for any $(k, q) \in \mathcal{S}$ and any $\varphi, \psi \in \mathcal{D}_+(T)$, the following inequality holds:

$$\begin{aligned} &\int_0^T\int_{\mathcal{R}} [|u^\epsilon - k|\varphi_t + |f(u^\epsilon) - f(k)|\varphi_x + |v^\epsilon - q|\psi_t]dx\ dt \\ &\qquad + \int_{\mathcal{R}} [|u^{\epsilon,0} - k|\varphi(x,0) + |v^{\epsilon,0} - q|\psi(x,0)]dx \\ &\qquad - \int_{\mathcal{R}} [|u^\epsilon(x,T) - k|\varphi(x,T) + |v^\epsilon(x,T) - q|\psi(x,T)]dx \\ &\quad \geq \frac{1}{\delta}\int_0^T\int_{\mathcal{R}} (\sigma(u^\epsilon - k)\varphi - \sigma(v^\epsilon - q)\psi)(A(u^\epsilon) - v^\epsilon)dx\ dt \\ &\qquad + \epsilon\int_0^T\int_{\mathcal{R}} [\sigma(u^\epsilon - k)\varphi_x u^\epsilon_x + \sigma(v^\epsilon - q)\psi_x v^\epsilon_x]dx\ dt. \end{aligned} \tag{6.7}$$

This variational inequality should be compared with the characterization (2.5) of the entropy solution of (2.2).

Now let $k = u(y,\tau),\ q = v(y,\tau)$, and $\varphi(x,t) = \psi(x,t) = \omega_\theta(x-y)\omega_\theta(t-\tau)$ in (6.7), where $(y,\tau) \in \mathcal{R} \times (0,T)$ is fixed, and integrate the result over $\mathcal{R} \times (0,T)$ with respect to y and τ. Consider (2.5) similarly, but use (y,τ) as integration variables instead of x and t. Then take $k = u^\epsilon(x,t),\ q = v^\epsilon(y,\tau)$, and

$$\varphi(y,\tau) = \psi(y,\tau) = \omega_\theta(x-y)\omega_\theta(t-\tau).$$

If we add the two inequalities, we then obtain

$$L(\epsilon,\theta) \leq R(\epsilon,\theta) - \frac{1}{\delta}I_1(\epsilon,\theta) - \epsilon I_2(\epsilon,\theta). \tag{6.8}$$

Here

$$L(\epsilon,\theta) = \int_0^T\int_{\mathcal{R}}\int_{\mathcal{R}} (|u^\epsilon(x,T) - u| + |v^\epsilon(x,T) - v|)\omega_\theta(x-y)\omega_\theta(T-\tau)dx\ dy\ d\tau$$

$$
\begin{aligned}
&+\int_0^T\int_{\mathcal{R}}\int_{\mathcal{R}}(|u^\epsilon-u(y,T)|+|v^\epsilon-v(y,T)|)\omega_\theta(x-y)\omega_\theta(T-t)dx\;dy\;dt,\\
R(\epsilon,\theta)=&\int_0^T\int_{\mathcal{R}}\int_{\mathcal{R}}(|u^{\epsilon,0}(x)-u|+|v^{\epsilon,0}(x)-v|)\omega_\theta(x-y)\omega_\theta(T)dx\;dy\;d\tau\\
&+\int_0^T\int_{\mathcal{R}}\int_{\mathcal{R}}(|u^\epsilon-u^0(y)|+|v^\epsilon-v^0(y)|)\omega_\theta(x-y)\omega_\theta(t)dx\;dy\;dt,\\
I_1(\epsilon,\theta)=&\int_0^T\int_{\mathcal{R}}\int_0^T\int_{\mathcal{R}}[(\sigma(u^\epsilon-u))(A(u^\epsilon)-A(u))+|v^\epsilon-v|]dx\;dt\;dy\;d\tau,\\
I_2(\epsilon,\theta)=&\int_0^T\int_{\mathcal{R}}\int_0^T\int_{\mathcal{R}}[(\sigma(u^\epsilon-u)u^\epsilon_x+\sigma(v^\epsilon-v)v^\epsilon_x)\omega_\theta{}'(x-y)\omega_\theta(t-\tau)]dx\;dt\;dy\;d\tau,
\end{aligned}
$$

where $u^\epsilon=u^\epsilon(x,t)$, $v^\epsilon=v^\epsilon(x,t)$, $u=u(y,\tau)$, and $v=v(y,\tau)$.

From the total-variation bounds and the L^1-continuity properties of the solutions (u^ϵ,v^ϵ) and (u,v), we obtain by a standard argument (cf., e.g., [26]) that there is a constant M, independent of ϵ, δ, and θ, such that

$$
\begin{aligned}
&|L(\epsilon,\theta)-\|u^\epsilon(\cdot,T)-u(\cdot,T)\|_1+\|v^\epsilon(\cdot,T)-v(\cdot,T)\|_1|\le M\theta,\\
&|R(\epsilon,\theta)-\|u^{\epsilon,0}-u^0\|_1+\|v^{\epsilon,0}-v^0\|_1|\le M\theta.
\end{aligned}
\tag{6.9}
$$

Furthermore, the monotonicity property of A implies that $I_1(\epsilon,\theta)\ge 0$. Finally, from the properties of (u^ϵ,v^ϵ), we obtain that

$$
|I_2(\epsilon,\theta)|\le T\sup_{0\le t\le T}(\|u^\epsilon_x(\cdot,t)\|_1+\|v^\epsilon_x(\cdot,t)\|_1)\int_{\mathcal{R}}|\omega_\theta{}'(y)|dy\le\frac{M}{\theta}.
$$

Hence, together with (6.2), (6.4), (6.8), and (6.9), this implies that

$$
\|u^\epsilon(\cdot,T)-u(\cdot,T)\|_1+\|v^\epsilon(\cdot,T)-v(\cdot,T)\|_1\le M\left(\epsilon+\theta+\frac{\epsilon}{\theta}\right).
$$

The desired result now follows by choosing $\theta=\epsilon^{1/2}$. □

In addition to the system studied above, we shall also need a regularized version of the equilibrium model. Consider the pure initial-value problem

$$
\begin{aligned}
(w^\epsilon+A(w^\epsilon))_t+f(w^\epsilon)_x&=\epsilon(w^\epsilon+A(w^\epsilon))_{xx},\\
w^\epsilon(x,0)&=u^{\epsilon,0}(x),
\end{aligned}
\tag{6.10}
$$

where $\{u^{\epsilon,0}\}\subset H^\infty(\mathcal{R})$ denotes the initial functions introduced at the beginning of this section. From the properties of $\{u^{\epsilon,0}\}$ and u^0 given above, it follows from the standard theory for scalar equations (cf., e.g., [10]) that for any $T>0$, there is a constant M, independent of ϵ, such that

$$
\|w(\cdot,t)-w^\epsilon(\cdot,t)\|_1\le M\epsilon^{1/2}\quad\text{for } 0\le t\le T.
\tag{6.11}
$$

Hence, if we write the error $u-w$ in the form

$$
u-w=(u-u^\epsilon)+(u^\epsilon-w^\epsilon)+(w^\epsilon-w),
$$

it remains to estimate the term $u^\epsilon-w^\epsilon$. In order to estimate this final error term, we shall rely on properties of system (5.1) derived in the previous section and on well-known properties of scalar diffusion equations of the form of (6.10). Consider a scalar equation of the form

$$\begin{aligned}(z + A(z))_t + f(z)_x &= \epsilon(z + A(z))_{xx} + G, \\ z(x, 0) &= z^0(x),\end{aligned} \tag{6.12}$$

where $z^0 = z^0(x)$ represents the initial function and $G = G(x, t)$ is a given source term. It is well known that if the data z^0 and G are smooth, then problem (6.12) has a unique smooth solution. Furthermore, if z and $\hat{z}$ denote two solutions of (6.12) with data (z^0, G) and $(\hat{z}^0, \hat{G})$, respectively, then we have the estimate

$$\|z(\cdot, t) - \hat{z}(\cdot, t)\|_1 \leq M_A \left(\|z^0 - \hat{z}^0\|_1 + \int_0^t \|G - \hat{G}\|_1 ds \right), \tag{6.13}$$

where the constant M_A only depends on the function A. In fact, this estimate can easily be proved by using arguments similar to those used in the proofs of Lemmas 5.2 and 5.3. Using this property of equation (6.12) and the properties of system (5.1) derived in section 5, we now establish the following estimate.

LEMMA 6.2. *Let $\{u^{\epsilon,0}\}$ and $\{v^{\epsilon,0}\}$ satisfy the requirements of Lemma* 6.1. *Then for any $T > 0$, there is a constant M, independent of ϵ and δ, such that*

$$\|u^\epsilon(\cdot, t) - w^\epsilon(\cdot, t)\|_1 \leq \frac{M\delta}{\epsilon} \quad \textit{for } 0 \leq t \leq T.$$

Proof. First, we consider system (5.1), and as above, we let

$$p^\epsilon = A(u^\epsilon) - v^\epsilon.$$

By adding the two equations of (5.1), we obtain

$$(u^\epsilon + v^\epsilon)_t + f(u^\epsilon)_x = \epsilon(u^\epsilon + v^\epsilon)_{xx}$$

or

$$(u^\epsilon + A(u^\epsilon) - p^\epsilon)_t + f(u^\epsilon)_x = \epsilon(u^\epsilon + A(u^\epsilon) - p^\epsilon)_{xx}. \tag{6.14}$$

Introduce a new function $z^\epsilon(x, t)$ defined by

$$z^\epsilon + A(z^\epsilon) = u^\epsilon + A(u^\epsilon) - p^\epsilon. \tag{6.15}$$

Then the differential equation (6.14) can be written in the form of (6.12), i.e.,

$$(z^\epsilon + A(z^\epsilon))_t + f(z^\epsilon)_x = \epsilon(z^\epsilon + A(z^\epsilon))_{xx} + G, \tag{6.16}$$

where

$$G = f(z^\epsilon)_x - f(u^\epsilon)_x.$$

From the monotonicity property of A, it follows that

$$|z^\epsilon(x, t) - u^\epsilon(x, t)| \leq |p^\epsilon(x, t)|.$$

Hence it follows from (6.2) and Lemma 5.3 that

$$\|z^\epsilon(\cdot, t) - u^\epsilon(\cdot, t)\|_1 \leq M\delta, \quad 0 \leq t \leq T. \tag{6.17}$$

By differentiating (6.15) with respect to x, we obtain

$$z_x^\epsilon - u_x^\epsilon + A'(z_x^\epsilon)(z_x^\epsilon - u_x^\epsilon) = (A'(u^\epsilon) - A'(z^\epsilon))u_x^\epsilon - p_x^\epsilon,$$

and this implies that

$$\|z_x^\epsilon(\cdot,t) - u_x^\epsilon(\cdot,t)\|_1 \leq M_A \|u^\epsilon(\cdot,t) - z^\epsilon(\cdot,t)\|_1 \|u_x^\epsilon(\cdot,t)\|_\infty + \|p_x^\epsilon(\cdot,t)\|_1.$$

From estimates (5.10), (5.13), (6.2), and (6.17), we therefore have

$$\|z_x^\epsilon(\cdot,t) - u_x^\epsilon(\cdot,t)\|_1 \leq \frac{M\delta}{\epsilon} \quad \text{for } 0 \leq t \leq T, \tag{6.18}$$

where M is independent of ϵ and δ. Now observe that the source term G introduced above can be represented in the form

$$G = f'(z^\epsilon)(z_x^\epsilon - u_x^\epsilon) + (f'(z^\epsilon) - f'(u^\epsilon))u_x^\epsilon,$$

and since $\|u_x^\epsilon(\cdot,t)\|_1$ is uniformly bounded on $[0,T]$ and since

$$\|z^\epsilon(\cdot,t) - u^\epsilon(\cdot,t)\|_\infty \leq \|z_x^\epsilon(\cdot,t) - u_x^\epsilon(\cdot,t)\|_1,$$

it follows from (6.18) that

$$\|G(\cdot,t)\|_1 \leq \frac{M\delta}{\epsilon} \quad \text{for } 0 \leq t \leq T. \tag{6.19}$$

Hence the stability estimate (6.13) implies that for $0 \leq t \leq T$,

$$\|z^\epsilon(\cdot,t) - w^\epsilon(\cdot,t)\|_1 \leq M\left(\|z^\epsilon(\cdot,0) - u^{\epsilon,0}\|_1 + \frac{\delta}{\epsilon}\right) \leq \frac{M\delta}{\epsilon}. \tag{6.20}$$

The desired estimate now follows from (6.17), (6.20), and the triangle inequality. □

Proof of Theorem 2.2. We can now easily complete the proof of Theorem 2.2. By writing

$$u - w = (u - u^\epsilon) + (u^\epsilon - w^\epsilon) + (w^\epsilon - w),$$

we obtain from Lemmas 6.1 and 6.2 and estimate (6.11) that for $0 \leq t \leq T$,

$$\|u(\cdot,t) - w(\cdot,t)\|_1 \leq M\left(\epsilon^{1/2} + \frac{\delta}{\epsilon}\right).$$

By choosing $\epsilon = \delta^{2/3}$, the desired estimate follows. □

REFERENCES

[1] J. Bear and Y. Bachmat, *Introduction to modeling of transport phenomena in porous media*, in Theory and Applications of Transport in Porous Media, Vol. 4, Kluwer Academic Publishers, Norwell, MA, 1991.

[2] 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.

[3] 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. 789–830.

[4] 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.

[5] P. Colella, A. Majda, and V. Roytburd, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1059–1080.

[6] 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.

[7] B. Hanouzet and R. Natalini, *Weakly coupled systems of quasilinear hyperbolic equations*, Quaderno 28/1993, Istituto per le Applicazioni del Calcolo, Rome.

[8] H. O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes equations*, Academic Press, New York, 1989.

[9] S. N. Kruzkov, *First order quasi linear equations with several space variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[10] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first order quasi linear equation*, Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[11] R. J. LeVeque and J. Wang, *A linear hyperbolic system with stiff source terms*, in Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects (Proc. 4th International Conference on Hyperbolic Problems), Vol. 43, A. Donato and F. Oliveri, eds.,Vieweg and Sohn/Ballen Booksellers International, Hauppauge, NY, 1992, pp. 401–408.

[12] A. Levy, *On Majda's model for dynamic combustion*, Comm. Partial Differential Equations, 17 (1992), pp. 657–698.

[13] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[14] B. J. Lucier, *Error bounds for the methods of Glimm, Godunov, and LeVeque*, SIAM J. Numer. Anal., 22 (1985), pp. 1074–1081.

[15] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math, 40 (1981), pp. 70–93.

[16] O. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl. Ser. 2, 26 (1963), pp. 95–172.

[17] R. B. Pember, *Numerical methods for hyperbolic conservation laws with stiff relaxation* I: *Spurious solutions*, SIAM J. Appl. Math., 53 (1993), pp. 1293–1330.

[18] R. B. Pember, *Numerical methods for hyperbolic conservation laws with stiff relaxation* II: *Higher-order Godunov methods*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 824–859.

[19] H. K. Rhee, R. Aris, and N. R. Amundsen, *First-Order Partial Differential Equations: Theory and Application of Hyperbolic Systems of Quasi Linear Equations*, Vol. I, Prentice–Hall International Series, Prentice–Hall, Englewood Cliffs, NJ, 1986.

[20] H. K. Rhee, R. Aris, and N. R. Amundsen, *First-Order Partial Differential Equations: Theory and Application of Hyperbolic Systems of Quasi Linear Equations*, Vol. II, Prentice–Hall International Series, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[21] S. Schochet, *The instant-response limit in Witham's nonlinear traffic-flow model: Uniform well-posedness and global existence*, Asymptotic Anal., 1 (1988), pp. 263–282.

[22] H. J. Schroll, A. Tveito, and R. Winther, *An L^1-error estimate for a semi-implicit difference scheme applied to a stiff system of conservation laws*, SIAM J. Numer. Anal., 34 (1997), to appear.

[23] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

[24] K. S. Sorbie, M. D. Yuan, A. C. Todd, and R. M. S. Wat, *The Modeling and Design of Scale Inhibitor Squeeze Treatments in Complex Reservoirs*, Society of Petroleum Engineers, Richardson, TX, 1991.

[25] Z. Teng and L. Ying, *Existence, uniqueness and convergence as vanishing viscosity for a reaction-diffusion-convection system*, Acta Math. Sinica (N.S.), 5 (1989), pp. 114–135.

[26] A. Tveito and R. Winther, *An error estimate for a finite difference scheme approximating a hyperbolic system of conservation laws*, SIAM J. Numer. Anal., 30 (1993), pp. 401–424.

[27] L. Ying and Z. Teng, *Riemann problem for a reacting and convection hyperbolic system*, Approx. Theory Appl., 1 (1984), pp. 95–122.

[28] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley, New York, 1973.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 162–177, January 1997

008

ISOCHRONOUS CENTERS IN PLANAR POLYNOMIAL SYSTEMS[*]

C. J. CHRISTOPHER[†] AND J. DEVLIN[‡]

Abstract. Two algorithms are given for finding conditions for a critical point to be an isochronous center. The first is based on a systematic search for a transformation to the simple harmonic oscillator and as an example is used to find conditions for an isochronous center in the Kukles system; the second algorithm is specific to systems with homogeneous nonlinearities and is based on a connection with an Abel differential equation. General properties of systems with isochronous centers are also considered and results on Liénard and Hamiltonian systems are deduced; a close connection is demonstrated between isochronous centers and complex centers.

Key words. polynomial systems, centers, isochronous centers

AMS subject classification. 34C25

PII. S0036141093259245

1. Introduction.

Consider the planar system

$$\dot{x} = P(x, y), \qquad \dot{y} = Q(x, y), \tag{1.1}$$

where P and Q are polynomials in x and y and the dot denotes differentiation with respect to time. Suppose that (1.1) has a center and define a period function T by taking a semitransversal through the center with some parametrization $\mathbf{x}(s)$; $T(s)$ is then the time taken for the trajectory starting at $\mathbf{x}(s)$ to return to $\mathbf{x}(s)$.

We are interested in the case when the period function is constant; the center is then said to be *isochronous.* Probably the first nonlinear example of isochronous oscillations is Huygen's pendulum, in which the bob is constrained to follow a cycloidal path; the period of the motion is then independent of the amplitude [12]. Recently, Needham [18] has demonstrated the existence of an isochronous center in a system arising in telecommunications theory—it is of interest to note that this system hence has the property that small changes in the initial conditions can alter the amplitude of an oscillation but not its frequency.

Questions relating to the period function have been studied by a number of authors. One motivation is a connection between solutions of certain second-order boundary value problems and properties of the period function for related systems (1.1): the range of values of the period determines the range of initial conditions for which the second-order problem has a solution [5, 21]. Also, in the study of subharmonic bifurcations of periodically forced Hamiltonian systems, a nondegeneracy condition is that the period function of the unperturbed system be locally strictly monotone [8]—monotonicity conditions have hence been of special interest [4]. In particular, quadratic Hamiltonian systems and Lotka–Volterra systems are both known to have monotone period functions [10, 23]. We note that in both these applications, isochronicity represents the worst possible case.

[*] Received by the editors November 30, 1993; accepted for publication (in revised form) October 13, 1995. This research was supported by the Science and Engineering Research Council.
http://www.siam.org/journals/sima/28-1/25924.html

[†] Department of Mathematics, The University of Wales, Aberystwyth, Dyfed SY23 3BZ, United Kingdom. Current address: School of Mathematics and Statistics, Plymouth University, Plymouth, Devon PL4 8AA, United Kingdom (cchristopher@plymouth.ac.uk).

[‡] Department of Mathematics, The University of Wales, Aberystwyth, Dyfed SY23 3BZ, United Kingdom (jdz@aber.ac.uk).

To fully understand the properties of the period function in a class of systems, it is necessary to consider the bifurcation of critical points of the period. This question has been studied by Chicone and Jacobs [6]. Suppose that the system

$$\dot{x} = -y + P_\lambda(x, y), \qquad \dot{y} = x + Q_\lambda(x, y)$$

has a center at the origin for all λ. If $T(x; \lambda)$ is the period of the solution through the point $(x, 0)$, then T is written as

$$T(x; \lambda) = 2\pi + a_0(\lambda)x + a_1(\lambda)x^2 + \cdots.$$

If $a_i(\lambda^*) = 0$ for $i = 0, \ldots, n$ and $a_{n+1}(\lambda^*) \neq 0$, then it is shown that at most n zeros of T_x bifurcate out of the origin of the system with $\lambda = \lambda^*$ when λ is perturbed. However, if $a_i(\lambda^*) = 0$ for all i so that the origin is isochronous, then it is necessary to determine the ideal generated by the $a_i(\lambda)$; this is obviously a much more difficult problem. Thus a complete analysis of these bifurcations for a class of systems requires the identification of the subclass of systems with an isochronous center.

Isochronous centers also play an important part in another paper of Chicone and Jacobs [7]. In this paper, they look at the bifurcation of limit cycles from periodic orbits around a center. Since the method used requires knowledge of the period function, it is natural to first investigate bifurcation from isochronous centers; it is shown that at most three limit cycles can bifurcate from a quadratic isochronous center.

Necessary and sufficient conditions for quadratic systems to have an isochronous center at the origin were found by Loud [14], and conditions for systems in which P and Q are cubic polynomials without quadratic terms were obtained by Pleshkan [19]. The system $d^2x/dy^2 + g(x) = 0$ was considered by Urabe [22]: the only case which yields an isochronous center is the simple harmonic oscillator $g(x) = k^2x$. Apart from these cases, knowledge of polynomial systems with isochronous centers is slight.

In this paper, we first consider some general properties of polynomial systems with isochronous centers; these allow us to prove, for example, that a polynomial Liénard system with an isochronous center has no other critical point. We then go on to present an algorithm for obtaining necessary conditions for a point to be an isochronous center, based on a systematic search for a transformation to the simple harmonic oscillator. This is similar to the Liapunov function method which has been applied with much success to the problem of determining necessary conditions for a center [3]. This approach to isochronous centers seems to have been used first by Pleshkan in his work on cubic systems [19]. We illustrate our method by applying it to the system

$$\dot{x} = y, \qquad \dot{y} = -x + ax^2 + bxy + cy^2 + dx^3 + ex^2y + fxy^2 + gy^3 \tag{1.2}$$

and obtaining necessary conditions for the origin to be an isochronous center. To complement this work on necessary conditions, we give some simple sufficient conditions for the existence of an isochronous center. We show that the conditions obtained for (1.2) are sufficient, thus completely solving the problem for this system. We end by examining systems with homogeneous nonlinearities in some detail; we give an alternative algorithm for this special case based on a transformation to an Abel equation and we give a new class of systems with isochronous centers.

2. Some general results. In this section, we present several results of a general nature. We first give a definition of what we mean by a center.

Definition 2.1. *We say that a critical point of system* (1.1) *is a* center *if there is a deleted neighborhood of the point which consists entirely of closed trajectories surrounding that point. The center is said to be* nondegenerate *if the linearized vector field at the point has two nonzero eigenvalues and* isochronous *if its associated period function is constant.*

Remark. It can be shown [20] that by an affine transformation, a polynomial system with a nondegenerate center can be brought to the form

$$\dot{x} = -\lambda y + p(x,y), \qquad \dot{y} = \lambda x + q(x,y), \tag{2.1}$$

where p and q are polynomials with all terms of degree at least two and $\lambda \neq 0$.

Theorem 2.2. *An isochronous center is nondegenerate.*

Proof. Suppose that the origin of (1.1) is a degenerate center. If the linear terms vanish, we have

$$\dot{r} = f(\theta)r^2 + O(r^3), \qquad \dot{\theta} = g(\theta)r + O(r^2),$$

where f and g are homogeneous cubic polynomials in $\cos\theta$ and $\sin\theta$. Then for any $\delta > 0$, there are trajectories sufficiently close to the origin along which $|\dot{\theta}| < \delta$. Thus there are periodic solutions of arbitrarily large period.

Now suppose that the system has degenerate linear terms. If the flow is not to be area contracting or expanding, then the divergence of the vector field must vanish at the critical point. Thus both eigenvalues vanish and an affine transformation brings the system to the form

$$\dot{x} = ky + O(x^2+y^2), \qquad \dot{y} = O(x^2+y^2).$$

Furthermore, since scaling time by a constant will have no effect on isochronicity, we can take $k = -1$. Thus, in polar coordinates,

$$\dot{r} = r\cos\theta\sin\theta + O(r^2), \qquad \dot{\theta} = \sin^2\theta + O(r).$$

Given $\delta > 0$, there exist trajectories sufficiently close to the origin along which $\dot{\theta} < \sin^2\theta + \delta^2$. Near the origin along the positive y-axis, trajectories cross into the first quadrant. Thus trajectories go around the origin in an anticlockwise direction. Ignoring any contribution to the period by parts of the trajectory where $\dot{\theta} < 0$, we find that the period is at least

$$\int_0^{2\pi} \frac{d\theta}{\sin^2\theta + \delta^2} = \frac{2\pi}{\delta(1+\delta^2)^{1/2}}.$$

Thus we can choose δ small enough to give a contradiction.

Theorem 2.3. *An isochronous period annulus cannot have any finite critical point on its boundary except for an isochronous center.*

Proof. Suppose that we have a critical point on the boundary of an isochronous period annulus D of period T; then given a neighborhood U of the point, there is a smaller neighborhood V such that any trajectory in V will take at least time $2T$ to pass out of U. Hence any trajectory passing through $V \cap D$ will remain in $U \cap D$ for all time. Since these trajectories are closed curves, the critical point must be a center.

Corollary 2.4. *If D is an isochronous period annulus, then any trajectory in $\partial\overline{D}$ is unbounded as t increases and as t decreases.*

Although these results might appear simple, they can reveal a surprising amount of information. For example, if the polynomial Liénard system

$$\dot{x} = y - f(x), \qquad \dot{y} = -g(x)$$

has an isochronous center, then either $f \equiv 0$ or g is of odd degree and there are no other critical points. (In fact, if $f \equiv 0$ then $g = k^2 x$ by the result of Urabe [22].) This is a consequence of the following more general result.

Theorem 2.5. *Suppose that*

(1) *$Q(x,y) = q(x)$, where q has more than one zero* (*counting multiplicity*);

(2) *if the zeros of $q(x)$ are x_i and $P(x,y)$ is of degree n in y, then the coefficient of y^n in $P(x,y)$ is never zero for* $\min\{x_i\} \le x \le \max\{x_i\}$;

(3) *if $P(x,y)$ is of degree m in x, then the coefficient of x^m in $P(x,y)$ is never zero.*

Then system (1.1) *cannot have an isochronous center.*

Proof. Suppose that there is an isochronous center at the point (x_0, y_0) so that $q(x_0) = 0$. Since the linearization of (1.1) at this point is

$$\dot{x} = P_x(x_0,y_0)x + P_y(x_0,y_0)y, \qquad \dot{y} = q'(x_0)x,$$

we have $q'(x_0) \neq 0$ by Theorem 2.2. Hence condition (1) implies the existence of $x_1 \neq x_0$ such that $q(x_1) = 0$.

Let the period annulus of the isochronous center be D. If there is a trajectory in $\overline{D}$ which intersects the line $x = x_1$, then there is a trajectory in $\overline{D}$ which touches the line at a point where $\dot{x} = 0$. But $\dot{y} = 0$ on $x = x_1$, so $\overline{D}$ contains a critical point other than (x_0, y_0), contradicting Theorem 2.3. Hence $\overline{D}$ cannot intersect the line $x = x_1$ for any $x_1 \neq x_0$ such that $q(x_1) = 0$.

Now suppose that there exist x_1 and x_2 with $q(x_1) = q(x_2) = 0$ and $x_1 < x_0 < x_2$. Then $\overline{D}$ is entirely contained in the strip $(x_1, x_2) \times \mathbf{R}$. Now by (1) and (2), there exists $M > 0$ such that

$$\left|\frac{dy}{dx}\right| = \left|\frac{q(x)}{P(x,y)}\right|$$

is bounded in $[x_1, x_2] \times \{(-\infty, -M) \cup (M, \infty)\}$, so no solution can become unbounded in the strip $(x_1, x_2) \times \mathbf{R}$. But $\overline{D}$ is unbounded by Corollary 2.4—a contradiction. It follows that if the zeros of $q(x)$ are x_i, then either $x_0 = \max\{x_i\}$ or $x_0 = \min\{x_i\}$. We consider the case $x_0 = \max\{x_i\}$; the other case is similar.

Let $x_1 < x_0$ be such that $q(x_1) = 0$ and $q(x) \neq 0$ for $x \in (x_1, x_0)$. Then $\partial\overline{D}$ has points in the strip $(x_1, x_0) \times \mathbf{R}$. But $\partial\overline{D}$ is bounded in this strip, as above, and does not intersect the line $x = x_1$. So ∂D passes through points (x_0, y_-) and (x_0, y_+) where $y_- < y_0 < y_+$. Now $\dot{y}$ is single-signed for $x \in (x_1, x_0)$ and for $x > x_0$. So $\partial\overline{D} \subset (x_1, \infty) \times (y_-, y_+)$. Since D is unbounded, for arbitrarily large x there exists $y \in (y_-, y_+)$ such that $P(x,y) = 0$. But this contradicts (3), so we obtain the required result—(1.1) cannot have an isochronous center.

We also have the following result precluding the existence of an isochronous center.

Theorem 2.6. *Suppose that p and q are not both identically zero and that* (2.1) *has no critical points at infinity. Then the origin cannot be an isochronous center.*

Proof. If the origin is an isochronous center with period annulus D, then the trajectory through any finite point in $\partial\overline{D}$ is unbounded by Corollary 2.4. This is impossible if there are no critical points at infinity; so $\overline{D}$ fills the plane and for any

$R > 0$ there are periodic solutions entirely contained in the region $r > R$. But for any $M > 0$, $|\dot{\theta}| > M$ for r sufficiently large, and there are solutions of arbitrarily small period—a contradiction.

COROLLARY 2.7. *Suppose that the Hamiltonian system*

$$\dot{x} = -H_y, \qquad \dot{y} = H_x \tag{2.2}$$

of degree $n \geq 2$ has an isochronous center. If $H = H_0 + H_1 + \cdots + H_{n+1}$, where H_i is a homogeneous polynomial in x and y of degree i, then H_{n+1} has a repeated real linear factor.

Proof. Corollary 2.4 and Theorem 2.6 together imply that either the line at infinity is composed entirely of critical points or there exists a critical point at infinity with a hyperbolic sector. For system (2.2), the first of these is impossible, while a necessary condition for the second is that H_{n+1} have a repeated real linear factor.

We shall show later (Theorem 5.6) that if $H_3 \equiv \cdots \equiv H_n \equiv 0$, then the origin of (2.2) cannot be an isochronous center.

It sometimes helps to think of (1.1) as a complex system—we thus allow x and y to take complex values; however, we shall keep time t real, thus avoiding the difficulty of solution curves becoming solution surfaces. The resulting complex system will be denoted $(1.1)^*$.

DEFINITION 2.8. *We say that a critical point of the complex system* $(1.1)^*$ *is a* center *if there is a deleted neighborhood of the point in* $\mathbf{C}^2$ *which consists entirely of closed trajectories.*

THEOREM 2.9. *A critical point of the real system* (1.1) *is isochronous if and only if it is a center for system* $(1.1)^*$.

Proof. ($\Longrightarrow$) Let $(\phi(t;x,y), \psi(t;x,y))$ be the solution of $(1.1)^*$ passing through the point $(x,y) \in \mathbf{C}^2$ at time $t = 0$. It is well known that ϕ and ψ are holomorphic within their maximal interval of existence. Suppose that the origin is an isochronous center for (1.1) with period T and take a sufficiently small deleted neighborhood of the origin $A \subset \mathbf{C}^2$ so that for all $(x,y) \in A$, $\phi(t;x,y)$ and $\psi(t;x,y)$ exist for $0 \leq t \leq T$. Consider a point $(x,y) \in \mathbf{R}^2 \cap A$. Clearly, there is a product neighborhood of (x,y) in A; call this $U \times V$. First, fix $v \in V \cap \mathbf{R}$ and consider $\phi(T;u,v) - u$. This is a holomorphic function of u vanishing on the line $U \cap \mathbf{R}$ and hence identically zero on U. Now consider $\phi(T;u,v) - u$ with $u \in U$ fixed. This is a holomorphic function of v vanishing on the line $V \cap \mathbf{R}$ and hence identically zero on V. Hence $\phi(T;u,v) - u$ vanishes for all $(u,v) \in U \times V$. A similar result holds for $\psi(T,u,v) - v$. The result follows by analytic continuation throughout the deleted neighborhood.

($\Longleftarrow$) It is clear that (1.1) has a center if $(1.1)^*$ has one, so it only remains to prove isochronicity. Take $\phi(t;x,y)$, $\psi(t;x,y)$, U, and V as before. Fix $v \in V$ and consider the period $T(u)$ of the trajectory through (u,v) for $u \in U$. Since the function $T(u)$ is smooth, it must have level curves in U. (Otherwise, both partial derivatives vanish throughout U and the result holds trivially.) Take one of these curves, say $T(u) = c$; then the functions $\phi(c;u,v) - u$ and $\psi(c;u,v) - v$ vanish on this curve and hence on the whole of U. Thus the period function is constant for fixed v. Similarly, the period function is constant for fixed u and the result follows.

3. An algorithm. From (2.1) and Theorem 2.4, we can bring system (1.1) with an isochronous center to the form

$$\dot{x} = -y + O(x^2 + y^2), \qquad \dot{y} = x + O(x^2 + y^2) \tag{3.1}$$

by an affine transformation and a scaling of time by a constant. Furthermore, it can be shown [20] that there is an analytic change of coordinates $X = x + F(x, y)$, $Y = y + G(x, y)$, which brings (3.1) to the form

$$\dot{X} = -Yf(X^2 + Y^2), \qquad \dot{Y} = Xf(X^2 + Y^2),$$

where f is analytic and $f(0) = 1$. If the center is to be isochronous, then f must be constant [7]; so

$$\dot{X} = -Y, \qquad \dot{Y} = X. \tag{3.2}$$

Thus there is an analytic function $H(x, y)$ of the form $x + F(x, y)$ such that

$$\frac{d^2H}{dt^2} + H = 0. \tag{3.3}$$

Conversely, if there exists such a function $H(x, y)$ for a given system (3.1), then, taking $X = H$ and $Y = -dH/dt$, we obtain an analytic change of coordinates bringing (3.1) to the form (3.2). The search for such a function provides the basis for our algorithm.

From (3.2), we can find an analytic first integral of the system

$$\Phi(x, y) = X^2 + Y^2 = x^2 + y^2 + O(x^2 + y^2)$$

such that $d\Phi/dt = 0$; and it is easy to verify that if H is a solution of (3.3), then

$$\left(aH + b\frac{dH}{dt}\right)\Phi^n = (ax - by)(x^2 + y^2)^n + O((x^2 + y^2)^{n+1}) \tag{3.4}$$

is also a solution of (3.3) for any constants a and b and for any positive integer n. Hence if we write $H = \sum_k H_k$, where $\deg(H_k) = k$ and $H_1 = x$, then by the linearity of (3.3), we have two degrees of freedom in the choice of H_k for each odd $k \geq 3$.

We wish therefore to solve the following problem: find necessary and sufficient conditions for the existence of an analytic function $H(x, y) = x + O(x^2 + y^2)$ such that

$$P^2H_{xx} + 2PQH_{xy} + Q^2H_{yy} + (P_xP + P_yQ)H_x + (Q_xP + Q_yQ)H_y + H = 0. \tag{3.5}$$

We proceed by determining successive terms in the expansion of H. Suppose that the polynomials H_j have been found for $j \leq k - 1$. Equating the terms of degree k in (3.5), we have

$$y^2H_{kxx} - 2xyH_{kxy} + x^2H_{kyy} + (1 - k)H_k = G, \tag{3.6}$$

where G is a homogeneous polynomial of degree k whose coefficients are polynomials in the coefficients of $H_j (j \leq k - 1)$. Consider the linear operator

$$y^2\frac{\partial^2}{\partial x^2} - 2xy\frac{\partial^2}{\partial x \partial y} + x^2\frac{\partial^2}{\partial y^2} + (1 - k) \tag{3.7}$$

acting on the space of monomials of degree k. To find the rank of this operator, we consider (3.7) under the change of variables $z = x + iy$, $\overline{z} = x - iy$, which gives

$$-z^2\frac{\partial^2}{\partial z^2} + 2z\overline{z}\frac{\partial^2}{\partial z \partial \overline{z}} - \overline{z}^2\frac{\partial^2}{\partial \overline{z}^2} + (1 - k). \tag{3.8}$$

It is easy to see that (3.8) has eigenfunctions $z^r\overline{z}^s(r+s=k)$ with eigenvalues $1-(r-s)^2$. Hence for the real case (3.7), we have eigenfunctions $az^r\overline{z}^s+\overline{a}z^s\overline{z}^r$ with eigenvalues $1-(r-s)^2$. Thus when k is even, the operator is of full rank and there is a unique H_k which satisfies equation (3.6). When k is odd, the operator has nullity two and the kernel consists of functions of the form

$$(a+ib)z^{(k-1)/2}\overline{z}^{(k+1)/2}+(a-ib)z^{(k+1)/2}\overline{z}^{(k-1)/2}=2(ax+by)(x^2+y^2)^{(k-1)/2}.$$

This corresponds to the degrees of freedom (3.4) in our choice of H. Without loss of generality, therefore, we can specify that for odd k the coefficients of xy^{k-1} and y^k in H_k vanish. Hence in order to solve (3.6), we will have two consistency conditions imposed on the coefficients of G corresponding to the coefficients of $z(z\overline{z})^{(k-1)/2}$ and $\overline{z}(z\overline{z})^{(k-1)/2}$ vanishing. These will be our necessary conditions for the existence of an isochronous center. Thus solving equation (3.6) for each k, we obtain an infinite number of conditions.

Pleshkan [19] has shown that the infinite set of conditions produced is also sufficient, but in practice it is simpler to prove that the conditions obtained are sufficient by other means; some sufficient conditions are given in the next section. It is easy to see by induction that the conditions will be polynomial equations in the coefficients of P and Q, which can be reduced to a finite number by Hilbert's basis theorem.

On experimenting with different classes of systems, we found that the method seems to work most effectively when the center conditions are derived first and then assimilated into the algorithm. This is indeed the method of Pleshkan. However, we shall now give an example where the center conditions have to be derived alongside those for isochronicity since the complete center conditions are unknown at present.

The Kukles system

$$\text{(3.9)} \qquad \dot{x}=y, \qquad \dot{y}=-x+ax^2+bxy+cy^2+dx^3+ex^2y+fxy^2+gy^3$$

has been the subject of a number of recent investigations (see [9, 13, 16, 17], for example). Although the general problem of finding complete conditions for (3.9) to have a center at the origin remains unsolved, the extra conditions imposed by seeking an isochronous center allow the problem to be solved completely. We are also able to find necessary and sufficient conditions for the complex system (3.9)* to have an isochronous center at the origin.

The first six pairs of conditions for the origin of (3.9) to be an isochronous center were calculated as described above. The calculations were performed using the computer algebra system REDUCE. For an isochronous center, it is necessary that all twelve polynomials vanish. The zero-set of the polynomials can be obtained in its simplest form using REDUCE's Groebner basis procedure; however, to avoid unnecessary computation time, REDUCE was first used interactively to simplify the polynomials, as follows:

The first two conditions are

$$bc+ab+e+3g=0, \qquad 4c^2+10ac+10a^2+b^2+9d+3f=0.$$

Hence we can substitute for the variables e and f. One of the next pairs of conditions now gives

(3.10)

$$(17c^3+48ac^2+45a^2c+2b^2c+10a^3+ab^2)b+6(c-a)bd+(15c^2+42ac+15a^2+3b^2)g=0.$$

We first suppose that $(c-a)b \neq 0$. Since the transformation

$$x \mapsto \frac{x}{k}, \qquad y \mapsto \frac{y}{k} \tag{3.11}$$

preserves the form of the system, we can suppose that $c - a = 1$. If we also take $g = \lambda b$, then (3.10) gives

$$d = -\frac{[(72a^2 + 72a + 3b^2 + 15)\lambda + 120a^3 + 192a^2 + 99a + 3ab^2 + 2b^2 + 17]}{6}.$$

If a Groebner basis is now calculated for the remaining nine polynomials, we find that we must have $a = -1/2$ and $b^2 + 1 = 0$ for a common zero. This is obviously not a real solution; however, if we consider (3.9) as a complex equation with a real time parameter, we obtain a system of the following form after rescaling to compensate for the transformation (3.11):

$$\dot{x} = y, \qquad \dot{y} = -x + \mu(x \pm iy)^2 + \nu(x \pm iy)^3, \tag{3.12}$$

where μ and ν are complex coefficients.

We next consider the case $b = 0$. If a Groebner basis is now calculated for the remaining ten conditions (that is, including (3.10)), then a common zero only occurs when $a = c = 0$ and $d^2 + g^2 = 0$. The system is then of the form (3.12) with $\mu = 0$. Finally, when $a = c$ and $b \neq 0$, a Groebner basis calculation yields $a = c = 0$ and $d = -b^2/9$, and the system is

$$\dot{x} = y, \qquad \dot{y} = -x + bxy - \frac{b^2x^3}{9}. \tag{3.13}$$

This has an isochronous center at the origin since the system

$$\dot{X} = Y + \frac{b(X^2 - Y^2)}{6}, \qquad \dot{Y} = -X + \frac{bXY}{3},$$

which has an isochronous center at the origin by Theorem 4.4, can be transformed to (3.13) by the substitution $x = X$, $y = Y + b(X^2 - Y^2)/6$.

We have thus proved the following result.

THEOREM 3.1. *The real system* (3.9) *has an isochronous center at the origin if and only if it is of the form* (3.13).

The results on necessary conditions for an isochronous center which we have derived above can easily be carried over to the corresponding complex Kukles system $(3.9)^*$, the coefficients now being complex. We thus obtain systems (3.12) and $(3.13)^*$ as necessary for an isochronous center. System $(3.13)^*$ has an isochronous center as in Theorem 2.8. We will show in the next section that system (3.12) has an isochronous center. Thus we have the following result.

THEOREM 3.2. *The complex system* $(3.9)^*$ *has an isochronous center at the origin if and only if it is of the form* (3.12) *or* $(3.13)^*$.

4. Sufficient conditions for an isochronous center. The algorithm of the previous section provides necessary conditions for an isochronous center. In this section, we shall give some simple sufficient conditions.

Consider a polynomial system of the form

$$\dot{x} = -y + p(x,y), \qquad \dot{y} = x + q(x,y), \tag{4.1}$$

where p and q do not contain linear terms. We first look at the apparently trivial case

$$\dot{\theta} = \frac{d(\tan^{-1}(y/x))}{dt} = 1.$$

That is, $xq(x, y) - yp(x, y) = 0$. This will hold if and only if the system is of the form

$$\dot{x} = -y + xf(x, y), \qquad \dot{y} = x + yf(x, y) \tag{4.2}$$

for some polynomial f with $f(0, 0) = 0$. Clearly, if such a system has a center at the origin, then it is isochronous. It only remains therefore to find conditions for system (4.2) to have a center at the origin; however, this is a nontrivial problem.

In polar coordinates, we have

$$\frac{dr}{d\theta} = \sum_{i=1}^{m} r^{i+1} f_i(\cos\theta, \sin\theta), \tag{4.3}$$

where $f = \sum_i f_i$, $\deg(f_i) = i$. We can now use Devlin [11] to obtain necessary conditions for a center in terms of integrals of the functions f_i; for example, if $f_i \equiv 0$ $(i = 1, \dots, j)$, then

$$\int_0^{2\pi} f_{j+1} d\theta = 0, \qquad \int_0^{2\pi} f_{j+2} d\theta = 0.$$

We can also give the following sufficient condition for an isochronous center.

Theorem 4.1. *If there exist a 2π-periodic function $\sigma(\theta)$ and functions $A_i(\sigma)$ such that $f_i(\cos\theta, \sin\theta) = A_i(\sigma) d\sigma/d\theta$, then system* (4.2) *has an isochronous center at the origin.*

Proof. Let $R(t; R_0, t_0)$ be the solution of the initial value problem

$$\frac{dR}{dt} = \sum_{i=1}^{m} A_i(t) R^{i+1}, \qquad R(t_0) = R_0.$$

Then the solution of (4.2) with $r(0) = r_0$ is $r(\theta) = R(\sigma(\theta);\ r_0,\ \sigma(0))$; clearly, this solution is 2π-periodic for r_0 sufficiently small.

How comprehensive is this center condition for (4.2)? It is easily verified that Theorem 4.1 includes the well-known symmetry condition for a center. If f is quadratic, then it follows from the proof of Theorem 5.2 of [2] that the condition is necessary as well as sufficient. The condition is also necessary if f is homogeneous; this follows from the next result.

Corollary 4.2. *If f is homogeneous, then the origin is an isochronous center for system* (4.2) *if and only if*

$$\int_0^{2\pi} f(\cos\theta, \sin\theta) d\theta = 0.$$

Proof. The sufficiency follows from Theorem 4.1, taking $\sigma(\theta) = \int_0^\theta f(\cos s, \sin s) ds$. The necessity follows from the fact that, as above, $\int_0^{2\pi} f_{j+1}$ must vanish.

In general, however, the condition of Theorem 4.1 is not sufficient for the origin of (4.2) to be an isochronous center: in [1], there is given an equation of the form (4.3) with $m = 2$, $\deg(f_1) = 3$, and $\deg(f_2) = 6$ for which all solutions in a neighborhood of the origin are 2π-periodic but this condition is not satisfied. Setting $\xi = r^3$ in this example, we obtain a counterexample with the f_i of the correct degree.

We now consider systems of the form (1.1) satisfying the Cauchy–Riemann equations

$$P_x = Q_y, \qquad P_y = -Q_x. \tag{4.4}$$

Thus $P + iQ$ is an analytic function of the complex variable $z = x + iy$, and system (1.1) can be written in the form

$$\dot{z} = f(z). \tag{4.5}$$

We shall say that (1.1) is a *Cauchy–Riemann system.*

THEOREM 4.3. *System* (4.5) *has an isochronous center at* $z_0 = x_0+iy_0$ *if and only if* $f(z_0) = 0$ *and* $f'(z_0)$ *is nonzero and purely imaginary. The period is* $|2\pi/f'(z_0)|$.

Proof. There is a critical point at z_0 if and only if $f(z_0) = 0$. Suppose that we have an isochronous center at z_0. From Theorem 2.4, we must have $f'(z_0) \neq 0$. Furthermore, we also need $\mathrm{Re}(f'(z_0)) = 0$; otherwise, the divergence of the vector field at z_0 will not vanish.

Conversely, suppose that the above conditions are satisfied. Let z_1 be a point in a neighborhood of z_0. We define the contour integral

$$R(z) = \int_{z_1}^{z} \frac{dz}{f(z)}.$$

From Cauchy's integral formula, as z moves once around z_0, the value of $R(z)$ increases by $2\pi i/f'(z_0)$. Thus $H(z) = \exp(f'(z_0)R(z))$ is a well-defined holomorphic function in a neighborhood of z_0. But $d^2H(z)/dt^2 = f'(z_0)^2H(z)$ and $f'(z_0)^2$ is a negative real number. So $z \mapsto H(z)$ defines an analytic change of coordinates to the simple harmonic oscillator and z_0 is an isochronous center. The period follows from elementary considerations.

COROLLARY 4.4. *A sufficient condition for system* (1.1) *to have an isochronous center at a point* p *is that it satisfy the Cauchy–Riemann equations* (4.4) *and* $P(p) = Q(p) = P_x(p) = 0$, *and* $P_y(p) \neq 0$.

Remark. The relation of the Cauchy–Riemann equations to isochronicity seems to have been noticed first by Pleshkan [19].

COROLLARY 4.5. *Consider system* (4.1) *with* p *and* q *homogeneous polynomials of degree* n *satisfying the Cauchy–Riemann equations. Then the origin is an isochronous center of period* 2π. *There are exactly* $n-1$ *other critical points and these are all isochronous centers of period* $2\pi/(n-1)$.

Proof. From the hypothesis, the system can be written in the form

$$\dot{z} = f(z) = iz + az^n$$

for some $a \in \mathbf{C}\backslash\{0\}$. Since $f'(0) = i$, the origin is an isochronous center of period 2π. Apart from the origin, there are $n-1$ other critical points, corresponding to roots of the equation $z^{n-1} = -i/a$. At these points, $f'(z) = i + naz^{n-1} = -(n-1)i$. The result follows from Theorem 4.3.

We now use Theorem 4.3 to give a class of complex systems with isochronous centers at the origin; this class includes the complex Kukles system (3.12).

THEOREM 4.6. *The origin is an isochronous center for the complex system*

$$\dot{x} = y + p(x \pm iy), \qquad \dot{y} = -x + q(x \pm iy), \tag*{(4.6)$^{\pm}$}$$

where p *and* q *are polynomials with complex coefficients and with all terms of degree two or more.*

Proof. It suffices to show that the origin is an isochronous center for $(4.6)^+$: if we transform $(4.6)^-$ by $y \mapsto -y$ and $t \mapsto -t$, we obtain a system of the form $(4.6)^+$. Let

$$u = x + iy = [\mathrm{Re}(x) - \mathrm{Im}(y)] + i[\mathrm{Im}(x) + \mathrm{Re}(y)],$$

$$v = x - iy = [\mathrm{Re}(x) + \mathrm{Im}(y)] + i[\mathrm{Im}(x) - \mathrm{Re}(y)].$$

Then

$$\dot{u} = -iu + f(u), \qquad \dot{v} = iv + g(u), \tag{4.7}$$

where $f = p + iq$, $g = p - iq$. By Theorem 4.3, $u(t)$ is 2π-periodic for $|u(0)|$ sufficiently small. Hence we only need to prove that $v(t)$ is 2π-periodic. But from (4.7),

$$v(t) = e^{it}\left[v(0) + \int_0^t e^{-is} g(u(s))ds\right],$$

so it suffices to show that

$$\int_0^{2\pi} e^{-is} g(u(s))ds = 0. \tag{4.8}$$

For this, more information on the solutions u is required.

Following the proof of Theorem 4.3, we set $H(u) = u \exp \int_0^u \gamma(\xi) d\xi$, where

$$\frac{1}{(-iu + f(u))} = i\left(\frac{1}{u} + \gamma(u)\right).$$

Then $dH(u)/dt = -iH(u)$, so $H(u(t)) = H(u(0))e^{-it}$. But $H(u) = u + O(u^2)$ as $u \to 0$, so the inverse function H^{-1} exists in a neighborhood of the origin and $H^{-1}(\xi) = \xi + O(\xi^2)$ as $\xi \to 0$. Hence

$$u(t) = H^{-1}(H(u(0))e^{-it}) = H(u(0))e^{-it} + \sum_{j=2}^{\infty} h_j e^{-ijt}.$$

Since $g(u)$ is a polynomial, equation (4.8) follows for $|u(0)|$ sufficiently small.

5. Systems with homogeneous nonlinearities. We now consider systems of the form

$$\dot{x} = y + p(x, y), \qquad \dot{y} = -x + q(x, y), \tag{5.1}$$

where p and q are homogeneous polynomials of degree n. We shall derive an alternative algorithm to find conditions for the origin of this system to be an isochronous center. Expressing the system in polar coordinates, we have

$$\dot{r} = f(\theta)r^n, \qquad \dot{\theta} = -1 + g(\theta)r^{n-1}, \tag{5.2}$$

where f and g are homogenous polynomials of degree $n+1$ in $\cos\theta$ and $\sin\theta$. We now take

$$\rho = \frac{r^{n-1}}{1 - r^{n-1}g(\theta)} \tag{5.3}$$

to bring the system to the Abel equation

$$\frac{d\rho}{d\theta} = -(n-1)f(\theta)g(\theta)\rho^3 + (g'(\theta) - (n-1)f(\theta))\rho^2. \tag{5.4}$$

There is much literature on the relationship between equations (5.1) and (5.4); many of the results can be found in the survey paper [15]. We shall write $\rho(\theta; c)$ for the solution of (5.4) with starting point c at $\theta = 0$. Clearly, periodic solutions of (5.1) sufficiently close to the origin are transformed to 2π-periodic solutions of (5.4).

THEOREM 5.1. *System* (5.1) *has an isochronous center at the origin if and only if, for the corresponding system* (5.4),

$$\rho(2\pi; c) = c, \qquad \int_0^{2\pi} g(\theta)\rho(\theta; c) = 0 \tag{5.5}$$

for all sufficiently small $c > 0$.

Proof. The first condition is that for a center, so we need only consider the period. Since $\dot\theta = -1 + gr^{n-1} = -1/(1 + g\rho)$, the period of the solution of (5.2) which is transformed to $\rho(\theta; c)$ is

$$\int_0^{2\pi} (1 + g(\theta)\rho(\theta; c))d\theta = 2\pi + \int_0^{2\pi} g(\theta)\rho(\theta; c)d\theta.$$

The result follows.

As in [2, 11], we write $\rho(\theta; c) = \sum_{i=1}^{\infty} a_i(\theta)c^i$ with $a_1(0) = 1$ and $a_i(0) = 0$ for $i > 1$. Hence from Theorem 5.1, we have an isochronous center if and only if, for all $i \geq 1$,

$$a_i(2\pi) = a_i(0), \qquad \int_0^{2\pi} g(\theta)a_i(\theta)d\theta = 0. \tag{5.6}$$

A formula to calculate the coefficients $a_i(\theta)$ is known [11]: if l, r, and σ_i are operators $C[0, 2\pi] \to C[0, 2\pi]$ defined by

$$(lh)(\theta) = -(n-1)\int_0^{\theta} f(s)g(s)h(s)ds, \qquad (rh)(\theta) = \int_0^{\theta} [g'(s) - (n-1)f(s)]h(s)ds,$$
$$(\sigma_0 h)(\theta) \equiv 0, \qquad (\sigma_1 h)(\theta) = h(\theta),$$
$$\sigma_i = (i-2)\sigma_{i-2}l + (i-1)\sigma_{i-1}r, \quad i \geq 2,$$

then $a_i(\theta) = (\sigma_i 1)(\theta)$, where 1 is the constant function $1(\theta) \equiv 1$. This together with (5.6) provides an alternative algorithm to that given in section 3. As an example of this formulation, we give the first few necessary conditions explicitly.

COROLLARY 5.2. *If system* (5.1) *has an isochronous center at the origin, then the following conditions hold:*

(A)

$$\int_0^{2\pi} g(\theta)d\theta = 0;$$

(B)

$$\int_0^{2\pi} f(\theta)d\theta = 0;$$

(C)

$$\int_0^{2\pi} g\left[g-(n-1)\int_0^{\theta} f\right] d\theta = 0;$$

(D)

$$\int_0^{2\pi} f(\theta)g(\theta)d\theta = 0;$$

$$\text{(E)}\qquad (n-1)\int_0^{2\pi} g\left[\int_0^{\theta} fg\right] d\theta - \int_0^{2\pi} g\left[g-(n-1)\int_0^{\theta} f\right]^2 d\theta = 0;$$

(F)

$$\int_0^{2\pi} fg\left[g-(n-1)\int_0^{\theta} f\right] d\theta = 0;$$

(G)

$$\int_0^{2\pi} fg\left[g-(n-1)\int_0^{\theta} f\right]^2 d\theta = 0;$$

$$\text{(H)}\quad (n-1)\int_0^{2\pi} [g'-(n-1)f]\cdot\left[\int_0^{\theta} fg\right]^2 d\theta - 2\int_0^{2\pi} fg\left[g-(n-1)\int_0^{\theta} f\right]^3 d\theta = 0.$$

Condition (5.5) is sufficient as well as necessary, and we now give an example where it is used to prove isochronicity.

THEOREM 5.3. *Suppose that there exist a positive integer l with $2l \le n$ and polynomials $U(x,y)$ and $V(x,y)$ such that*

$$xp+yq=(x^2+y^2)^lU, \qquad xq-yp=(x^2+y^2)^lV,$$

$$(n-1)U_x=(n+1-2l)V_y, \qquad (n-1)U_y=-(n+1-2l)V_x.$$

Then the origin is an isochronous center for system (5.1).

Proof. Let $m=n+1-2l$. Then U and V are homogeneous polynomials of degree m and $(n-1)U+imV$ is a holomorphic function of the complex variable $z=x+iy$; so

$$(n-1)U+imV+(a+ib)z^m=r^m(m\sigma-i\sigma'),$$

where $\sigma(\theta)=(a\cos(m\theta)-b\sin(m\theta))/m$. Thus

$$f(\theta)=\frac{U}{r^m}=\frac{m\sigma}{(n-1)}, \qquad g(\theta)=\frac{V}{r^m}=-\frac{\sigma'}{m}, \tag{5.7}$$

and the Abel equation (5.4) reduces to $d\rho/d\theta=\rho^3\sigma\sigma'$. This has solution

$$\rho(\theta;c)=(c^{-2}+\sigma(0)^2-\sigma(\theta)^2)^{-1/2}$$

and since $g(\theta)=-\sigma'/m$, condition (5.5) holds.

Using (5.7), it is straightforward to verify that the conditions of the theorem are satisfied if and only if

$$\dot z=-iz+z^l\bar z^{l-1}\frac{[(n-l)(a+ib)z^m+(1-l)(a-ib)\bar z^m]}{m(n-1)}.$$

(Note that for $l = 1$ this reduces to a Cauchy–Riemann system.) For $z \in \mathbf{C}\backslash\{0\}$, we make the change of variables $w = z(z\bar{z})^{(l-1)/m}$. This gives the system

$$\dot{w} = -iw + \frac{(a+ib)w^{m+1}}{m};$$

by Corollary 4.5, all critical points of this system are isochronous centers. Hence we have the following result.

Corollary 5.4. *If the conditions of Theorem* 5.3 *are satisfied, then every critical point of the system is an isochronous center.*

A similar property holds for quadratic and cubic systems.

Theorem 5.5. *If a polynomial system with quadratic or cubic homogeneous nonlinearities has an isochronous center at the origin, then every critical point of the system is an isochronous center.*

Proof. For the quadratic case, the classification of isochronous centers is given by Loud [14]: the origin is an isochronous center if and only if the system can be brought to the form

$$\dot{x} = -y + Bxy, \qquad \dot{y} = x + Dx^2 + Fy^2$$

with $(D/B, F/B) = (-1/2, 1/2)$, $(0,1)$, $(0,1/4)$, or $(-1/2, 2)$. The first of these yields a Cauchy–Riemann system and so is covered by Corollary 4.5. The second and third have no other critical points, and the fourth has a symmetry $x \mapsto 2/B - x$ so that the other critical point $(2/B, 0)$ is also an isochronous center.

The cubic case was examined by Pleshkan [19], who gave the following condition: the origin is an isochronous center if and only if the system can be brought to one of the following forms:

$$(1) \quad \begin{cases} \dot{x} = -y - ax^3 - 3bx^2y + 3axy^2 + by^3, \\ \dot{y} = x + bx^3 - 3ax^2y - 3bxy^2 + ay^3, \end{cases}$$

$$(2) \quad \begin{cases} \dot{x} = -y - ax^3 + bx^2y + axy^2, \\ \dot{y} = x - ax^2y + bxy^2 + ay^3, \end{cases}$$

$$(3) \quad \begin{cases} \dot{x} = -y + 3ax^2y, \\ \dot{y} = x - 2ax^3 + 9axy^2. \end{cases}$$

The first of these is a Cauchy–Riemann system, while for the second $\dot{\theta} = 1$; the result thus holds for these two cases. In the third case, if $a \leq 0$ then there are no other critical points, so we take $a > 0$. Since the system is symmetric about the y-axis, we only need to consider the critical point at $((2a)^{-1/2}, 0)$. The change of coordinates

$$X = \frac{1}{2} - \left(\frac{a}{2}\right)^{1/2} x, \qquad Y = \frac{1}{2}\left(\frac{a}{2}\right)^{1/2} y$$

brings this point to the origin, and the system becomes

$$\dot{X} = -Y(1 - 12P), \qquad \dot{Y} = (P + 18Y^2)P',$$

where $P = X - X^2$. In a neighborhood of the origin, define

$$u = (1 - 12P)^{-3/2}PP', \qquad v = Y(1 - 12P)^{-3/2};$$

this brings the system to the form $\dot{u} = -v$, $\dot{v} = u$, so the critical point is indeed an isochronous center.

Finally, we consider Hamiltonian systems

$$\dot{x} = -H_y, \qquad \dot{y} = H_x \tag{5.8}$$

with homogeneous nonlinearities. If the origin is an isochronous center, then by Theorem 2.2 and by scaling time by a constant, we can suppose the Hamiltonian H to be of the form $H = (x^2 + y^2)/2 + h(x, y)$, where h is homogeneous. Under these conditions, we can improve on Corollary 2.7.

THEOREM 5.6. *The Hamiltonian system* (5.8) *with* $H = (x^2+y^2)/2+h(x, y)$ *has an isochronous center at the origin if and only if* $h \equiv 0$.

Proof. Clearly, if $h \equiv 0$, then the origin is an isochronous center. For the converse, we suppose that the origin is an isochronous center and that $h \not\equiv 0$ and consider trajectories in the period annulus, that is, the level curves $H = c$ as c increases from zero. By Corollary 2.4, there are level curves in the period annulus with points arbitrarily far from the origin. Write $H = r^2/2 + g(\theta)r^n$, where g is a homogeneous polynomial in $\cos\theta$ and $\sin\theta$ of degree n. The distance of the curve $H = c$ from the origin achieves its maximum on the ray $\theta = \theta_0$ where $g(\theta)$ achieves its minimum. By Corollary 5.2, $g(\theta)$ changes sign and so $g(\theta_0) < 0$. Hence there is a critical point on the ray $\theta = \theta_0$ at the point where

$$r = (-ng(\theta_0))^{-1/(n-2)}.$$

Thus it is not possible for there to be points in the period annulus arbitrarily far from the origin—a contradiction.

Acknowledgment. Both authors wish to thank the Science and Engineering Research Council for its funding of this research.

REFERENCES

[1] M. A. M. ALWASH, *On a condition for a centre of cubic non-autonomous systems*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 289–291.

[2] M. A. M. ALWASH AND N. G. LLOYD, *Non-autonomous equations related to polynomial two-dimensionals systems*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 129–152.

[3] T. R. BLOWS AND N. G. LLOYD, *The number of limit cycles of certain polynomial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 215–239.

[4] C. CHICONE, *The monotonicity of the period function for planar Hamiltonian vector fields*, J. Differential Equations, 69 (1987), pp. 310–321.

[5] C. CHICONE, *Geometric methods for two-point nonlinear boundary value problems*, J. Differential Equations, 72 (1988), pp. 360–407.

[6] C. CHICONE AND M. JACOBS, *Bifurcation of critical periods for plane vector fields*, Trans. Amer. Math. Soc., 312 (1989), pp. 433–486.

[7] C. CHICONE AND M. JACOBS, *Limit cycle bifurcations from quadratic isochrones*, J. Differential Equations, 91 (1991), pp. 268–327.

[8] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[9] C. CHRISTOPHER AND N. G. LLOYD, *On the paper of Jin and Wang concerning the conditions for a centre in certain cubic systems*, Bull. London Math. Soc., 22 (1990), pp. 5–12.

[10] W. A. COPPEL AND L. GAVRILOV, *The period of a Hamiltonian quadratic system*, Differential Integral Equations, 6 (1993), pp. 1357–1365.

[11] J. DEVLIN, *Word problems related to periodic solutions of a non-autonomous system*, Math. Proc. Cambridge Philos. Soc., 108 (1990), pp. 127–151.

[12] G. R. FOWLES AND G. L. CASSIDY, *Analytical Mechanics*, Saunders College Publishing, Philadelphia, Orlando, FL, 1993.

[13] X. Jin and D. Wang, *On the conditions of Kukles' for the existence of a centre*, Bull. London Math. Soc., 22 (1990), pp. 1–4.

[14] W. S. Loud, *Behaviour of the period of solutions of certain plane autonomous systems near centres*, J. Differential Equations, 3 (1964), pp. 21–36.

[15] N. G. Lloyd, *Limit cycles in polynomial systems: Some recent developments*, in New Directions in Dynamical Systems, T. Bedford and J. Swift, eds., London Mathematical Society Lecture Notes Series 127, Cambridge University Press, Cambridge, UK, 1988, pp. 653–669.

[16] N. G. Lloyd and J. M. Pearson, *Conditions for a centre and the bifurcation of limit cycles in a class of cubic systems*, in Bifurcations of Planar Vector Fields, P. Francoise and R. Roussarie, eds., Springer-Verlag, Berlin, New York, Heidelberg, 1990, pp. 230–242.

[17] N. G. Lloyd and J. M. Pearson, *Computing centre conditions for certain cubic systems*, J. Comput. Appl. Math., 40 (1991), pp. 323–336.

[18] D. J. Needham, *A centre theorem for two-dimensional complex holomorphic systems and its generalizations*, Proc. Roy. Soc. London Ser. A, 450 (1995), pp. 225–232.

[19] L. L. Pleshkan, *A new method of investigating the isochronicity of a system of two differential equations*, Differential Equations, 5 (1968), pp. 796–802.

[20] C. K. Seigel and J. K. Moser, *Lectures in Celestial Mechanics*, Springer-Verlag, New York, 1971.

[21] J. Smoller and A. Wasserman, *Global bifurcation of steady state solutions*, J. Differential Equations, 39 (1981), pp. 269–290.

[22] M. Urabe, *Potential forces which yield periodic motion of a fixed period*, J. Math Mech., 10 (1961), pp. 569–578.

[23] J. Waldvogel, *The period in the Lotka–Volterra system is monotonic*, J. Math. Anal. Appl., 114 (1986), pp. 178–184.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 178–188, January 1997

009

EFFECTIVE REDUCIBILITY OF QUASI-PERIODIC LINEAR EQUATIONS CLOSE TO CONSTANT COEFFICIENTS*

ÀNGEL JORBA†, RAFAEL RAMÍREZ-ROS†, AND JORDI VILLANUEVA†

Abstract. Let us consider the differential equation

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x, \quad |\varepsilon| \leq \varepsilon_0,$$

where A is an elliptic constant matrix and Q depends on time in a quasi-periodic (and analytic) way. It is also assumed that the eigenvalues of A and the basic frequencies of Q satisfy a diophantine condition. Then it is proved that this system can be reduced to

$$\dot{y} = (A^*(\varepsilon) + \varepsilon R^*(t, \varepsilon))y, \quad |\varepsilon| \leq \varepsilon_0,$$

where R^* is exponentially small in ε, and the linear change of variables that performs such a reduction is also quasi-periodic with the same basic frequencies as Q. The results are illustrated and discussed in a practical example.

Key words. quasi-periodic Floquet theorem, quasi-periodic perturbations, reducibility of linear equations

AMS subject classifications. 34A30, 34C20, 34C27, 34C50, 58F30

PII. S0036141095280967

1. Introduction.

The well-known Floquet theorem states that any linear periodic system $\dot{x} = A(t)x$ can be reduced to constant coefficients $\dot{y} = By$ by means of a periodic change of variables. Moreover, this change of variables can be taken over $\mathbb{C}$ with the same period as $A(t)$.

A natural extension is to consider the case in which the matrix $A(t)$ depends on time in a quasi-periodic way. Before beginning the discussion of this issue, let us recall the definition and basic properties of quasi-periodic functions.

DEFINITION 1.1. *A function f is a* quasi-periodic function *with a vector of basic frequencies $\omega = (\omega_1, \ldots, \omega_r)$ if $f(t) = F(\theta_1, \ldots, \theta_r)$, where F is 2π periodic in all its arguments and $\theta_j = \omega_j t$ for $j = 1, \ldots, r$. Moreover, f is called* analytic *on a strip of width ρ if F is analytical on an open set containing $|\operatorname{Im} \theta_j| \leq \rho$ for $j = 1, \ldots, r$.*

It is also known that an analytic quasi-periodic function $f(t)$ on a strip of width ρ has Fourier coefficients defined by

$$f_k = \frac{1}{(2\pi)^r} \int_{\mathbb{T}^r} F(\theta_1, \ldots, \theta_r) e^{-(k,\theta)\sqrt{-1}} \, d\theta$$

such that f can be expanded as

$$f(t) = \sum_{k \in \mathbb{Z}^r} f_k e^{(k,\omega)\sqrt{-1}t}$$

for all t such that $|\operatorname{Im} t| \leq \rho / \|\omega\|_\infty$. We denote by $\|f\|_\rho$ the norm

$$\|f\|_\rho = \sum_{k \in \mathbb{Z}^r} |f_k| e^{|k|\rho},$$

* Received by the editors February 1, 1995; accepted for publication (in revised form) October 27, 1995. This research was supported by UPC grant PR9409. The research of the first author was partially supported by DGICYT grant PB94-0215 (Spain), CIRIT grant GRQ93-1135 (Catalunya), and European Network grant ERB-CHRXCT 94060.

http://www.siam.org/journals/sima/28-1/28096.html

† Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (jorba@ma1.upc.es, rafael@tere.upc.es, jordi@tere.upc.es).

and it is not difficult to check that it is well defined for any analytical quasi-periodic function defined on a strip of width ρ. Finally, to define an analytic quasi-periodic matrix, we note that all of these definitions hold when f is a matrix-valued function. In this case, to define $\|f\|_\rho$, we use the infinity norm (which will be denoted by $|\cdot|_\infty$) for the matrices f_k.

With these definitions and properties, let us return to the problem of the reducibility of a linear quasi-periodic equation $\dot{x} = \widehat{A}(t)x$ to constant coefficients. The approach of this paper is to assume that the system is close to constant coefficients, that is, $\widehat{A}(t) = A + \varepsilon Q(t, \varepsilon)$, where ε is small. This case has already been considered in many papers (see [2], [8], and [9] among others), and the results can be summarized as follows. Let λ_i be the eigenvalues of A and let $\alpha_{ij} = \lambda_i - \lambda_j$ for $i \neq j$. Then if all the values $\operatorname{Re}\alpha_{ij}$ are different from zero, the reduction can be performed for $|\varepsilon| < \varepsilon_0$, ε_0 sufficiently small (see [2]). If some of the $\operatorname{Re}\alpha_{ij}$ are zero (this happens, for instance, if A is elliptic, that is, if all the λ_i are on the imaginary axis), more hypotheses are needed—usually these are (i) a diophantine condition involving the α_{ij} and the basic frequencies of $Q(t, \varepsilon)$ and (ii) to assume a nondegeneracy condition with respect to ε on the corresponding $\alpha_{ij}(\varepsilon)$ of the matrix $A + \varepsilon\overline{Q}(\varepsilon)$ ($\overline{Q}(\varepsilon)$ denotes the average of $Q(t, \varepsilon)$). This allows to prove (see [9] for details) that there exists a Cantorian set $\mathcal{E}$ such that the reduction can be performed for all $\varepsilon \in \mathcal{E}$. Moreover, the relative measure of the set $[0, \varepsilon_0] \setminus \mathcal{E}$ in $[0, \varepsilon_0]$ is exponentially small in ε_0.

Our purpose here is somewhat different. Instead of looking for a total reduction to constant coefficients (this seems to lead us to eliminate a dense set of values of ε; see [8] or [9]), we try to minimize the quasi-periodic part without taking out any value of ε. The result obtained is that the quasi-periodic part can be made exponentially small. Since all of the proof is constructive (and can be carried out with a finite number of steps), it can be applied to practical examples in order to perform an “effective” reduction: if ε is small enough, the remainder will be so small that, for practical purposes, it can be taken equal to zero. The error produced with this dropping can be easily bounded by means of the Gronwall lemma. Finally, we want to stress that we have also eliminated the nondegeneracy hypothesis of previous papers [8], [9].

Before finishing this introduction, we want to mention a similar result obtained when the dynamics of the system is slow: $\dot{x} = \varepsilon(A + \varepsilon Q(t, \varepsilon))x$. This case is contained in [14], which is an extension of [12]. The result obtained is also that the quasi-periodic part can be made exponentially small in ε. Total reducibility has been also considered in this case: in [15], it is stated that the reduction can be performed except for a set of values of ε of measure exponentially small.

There are many other results for the reducibility problem. For instance, in the case of the Schrödinger equation with quasi-periodic potential, we mention [3], [4], [5], [10], [11], and [13]. Another classical and remarkable paper is [7], where the general case (that is, without asking to be close to constant coefficients) is considered. Finally, classical results for quasi-periodic systems can be found in [6].

In order to simplify reading, the paper has been divided as follows. Section 2 contains the exposition (without technical details) of the main ideas and methodology, section 3 contains the main theorem, sections 4 and 5 are devoted to the proofs and, finally, section 6 contains an example to show how these results can be applied to a concrete problem.

2. The method. The method used is based on the same inductive scheme as [8]. Let us write our equation as

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x, \tag{1}$$

where A is an elliptic $d \times d$ matrix and $Q(t,\varepsilon)$ is quasi-periodic with $\omega = (\omega_1, \dots, \omega_r)$ as vector of basic frequencies and analytic on a strip of width ρ. First of all, let us rewrite this equation as

$$\dot{x} = (A_0(\varepsilon) + \varepsilon \widetilde{Q}(t,\varepsilon))x,$$

where $A_0(\varepsilon) = A + \overline{Q}(\varepsilon)$ and $\widetilde{Q}(t,\varepsilon) = Q(t,\varepsilon) - \overline{Q}(\varepsilon)$. Now let us assume that we are able to find a quasi-periodic $d \times d$ matrix P (with the same basic frequencies as Q) verifying

$$\dot{P} = A_0(\varepsilon)P - PA_0(\varepsilon) + \widetilde{Q}(t,\varepsilon) \tag{2}$$

such that $\|\varepsilon P(t,\varepsilon)\|_\sigma < 1$ for some $\sigma > 0$. In this case, it is not difficult to check that the change of variables $x = (I + \varepsilon P(t,\varepsilon))y$ transforms equation (1) into

$$\dot{y} = (A_0(\varepsilon) + \varepsilon^2 (I + \varepsilon P(t,\varepsilon))^{-1} \widetilde{Q}(t,\varepsilon) P(t,\varepsilon))y. \tag{3}$$

Since this equation is similar to (1) but with ε^2 instead of ε, the inductive scheme seems clear: average the quasi-periodic part of (3) and restart this process. The main difficulty that appears in this process comes from equation (2) because the solution contains the denominators $\lambda_i(\varepsilon) - \lambda_j(\varepsilon) + \sqrt{-1}(k,\omega)$, $1 \le i,\, j \le d$, where $\lambda_i(\varepsilon)$ are the eigenvalues of $A_0(\varepsilon)$. (This is shown in the proof of Lemma 4.2.) This divisor appears in the kth Fourier coefficient of P. Note that if the values $\lambda_i(\varepsilon) - \lambda_j(\varepsilon)$ are outside the imaginary axis, the (modulus of the) divisor can be bounded from below, making it easy to prove the convergence. On the other hand, the value $\lambda_i(\varepsilon) - \lambda_j(\varepsilon) + \sqrt{-1}(k,\omega)$ can be arbitrarily small, giving rise to convergence problems.

2.1. Avoiding the small divisors. Let us begin by assuming that the eigenvalues λ_i of the original unperturbed matrix A (see equation (1)) and the basic frequencies of Q satisfy the diophantine condition

$$|\lambda_i - \lambda_j + \sqrt{-1}(k,\omega)| \ge \frac{c}{|k|^\gamma} \quad \forall\, k \in \mathbb{Z}^r \setminus \{0\}, \tag{4}$$

where $|k| = |k_1| + \cdots + |k_r|$. Note that, in principle, we cannot guarantee that this condition holds in equation (2) because the eigenvalues of $A_0(\varepsilon)$ have been changed with respect to the ones of A (by an amount of $\mathcal{O}(\varepsilon)$) and some of the divisors can be very small or even zero.

The key point is to realize that as the eigenvalues of A move by an amount of $\mathcal{O}(\varepsilon)$ at most, the quantities $\lambda_i(\varepsilon) - \lambda_j(\varepsilon)$ are contained in a (complex) ball $B_{i,j}(\varepsilon)$ centered in $\lambda_i - \lambda_j$ and with radius $\mathcal{O}(\varepsilon)$. Since the center of the ball satisfies condition (4), the values (k,ω) cannot be inside that ball if $|k|$ is less than some value $M(\varepsilon)$. This implies that it is possible to cancel all of the harmonics such that $0 < |k| < M(\varepsilon)$ because they do not produce small divisors. (Note that we can have resonances only when (k,ω) is inside $B_{i,j}(\varepsilon)$.) The harmonics with $|k| \ge M(\varepsilon)$ are exponentially small in $M(\varepsilon)$ (when $M(\varepsilon) \to \infty$), that is, exponentially small in ε (when $\varepsilon \to 0$), so we do not need to eliminate them.

The idea of considering only frequencies less than some threshold M has already been applied in other contexts (see, for instance, [1]).

2.2. The iterative scheme. To apply the considerations above, we define, as before, $A_0(\varepsilon) = A + \varepsilon\overline{Q}(\varepsilon)$, $\widetilde{Q}(t,\varepsilon) = Q(t,\varepsilon) - \overline{Q}(\varepsilon)$ and we split $\widetilde{Q}(t,\varepsilon)$ into the sum of two matrices $Q_0(t,\varepsilon)$ and $R_0(t,\varepsilon)$: $Q_0(t,\varepsilon)$ contains the harmonics $Q_k e^{(k,\omega)\sqrt{-1}t}$

with $|k| < M(\varepsilon)$ and $R_0(t,\varepsilon)$ contains those with $|k| \geq M(\varepsilon)$. Therefore, (1) can be rewritten as

$$\dot{x} = (A_0(\varepsilon) + \varepsilon Q_0(t,\varepsilon) + \varepsilon R_0(t,\varepsilon))x. \tag{5}$$

Now the idea is to cancel $Q_0(t,\varepsilon)$ and to leave $R_0(t,\varepsilon)$. (It is already exponentially small with ε.) Therefore, we compute P_0 such that

$$\dot{P}_0 = A_0(\varepsilon)P_0 - P_0A_0(\varepsilon) + Q_0(t,\varepsilon).$$

Then the change $x = (I + \varepsilon P_0(t,\varepsilon))y$ gives

$$\dot{y} = \left[A_0 + \varepsilon^2(I+\varepsilon P_0)^{-1}Q_0P_0 + \varepsilon(I+\varepsilon P_0)^{-1}R_0(I+\varepsilon P_0)\right]y.$$

This equation can be rewritten to be like (5) to repeat the process. Note that the size of the harmonics with $0 < |k| < M(\varepsilon)$ has been squared. As we will see in the proofs, this is enough to guarantee convergence of those terms to zero. Thus the final equation has a purely quasi-periodic part that is exponentially small with ε.

2.3. Remarks. It is interesting to note that it is enough to apply a finite number of steps of the inductive process. We do not need to completely cancel the harmonics with $0 < |k| < M(\varepsilon)$, but we can stop the process when they are of the same size as those of R. (From the proof, it can be seen that the number of steps needed to achieve this is of order $|\ln|\varepsilon||$.) This allows us to (with the help of a computer) apply this procedure on a practical example.

Another remarkable point concerns the diophantine condition. Note that we need the condition only up to a finite order ($M(\varepsilon)$, which is of order $(1/|\varepsilon|)^{1/\gamma}$, as we shall see in the proofs). This means that in a practical example when the perturbing frequencies are known with finite precision, the diophantine condition can be easily checked.

3. The theorem. In what follows, $\mathcal{Q}_d(\rho,\omega)$ stands for the set of analytic quasi-periodic $d \times d$ matrices on a strip of width ρ and that have ω as their vector of basic frequencies. Moreover, i will denote $\sqrt{-1}$.

THEOREM 3.1. *Consider the equation $\dot{x} = (A + \varepsilon Q(t,\varepsilon))x$, $|\varepsilon| \leq \varepsilon_0$, $x \in \mathbb{R}^d$, where we have the following hypotheses:*

1. *A is a constant $d \times d$ matrix with different eigenvalues $\lambda_1, \dots, \lambda_d$.*
2. *$Q(\cdot,\varepsilon) \in \mathcal{Q}_d(\rho,\omega)$ with $\|Q(\cdot,\varepsilon)\|_\rho \leq q$ $\forall |\varepsilon| \leq \varepsilon_0$, for some $\omega \in \mathbb{R}^r$, and where $q, \rho > 0$.*
3. *The vector ω satisfies the diophantine conditions*

$$|\lambda_j - \lambda_\ell + i(k,\omega)| \geq \frac{c}{|k|^\gamma} \quad \forall\, k \in \mathbb{Z}^r \setminus \{0\} \quad \forall\, j,\, \ell \in \{1,\dots,d\} \tag{6}$$

for some constants $c > 0$ and $\gamma > r - 1$. As usual, $|k| = |k_1| + \cdots + |k_r|$.

Then there exist positive constants ε^*, a^*, r^*, and m such that for all ε, $|\varepsilon| \leq \varepsilon^*$, the initial equation can be transformed into

$$\dot{y} = (A^*(\varepsilon) + \varepsilon R^*(t,\varepsilon))y, \tag{7}$$

where

1. *A^* is a constant matrix with $|A^*(\varepsilon) - A|_\infty \leq a^*|\varepsilon|$ and*
2. *$R^*(\cdot,\varepsilon) \in \mathcal{Q}_d(\rho,\omega)$ and $\|R^*(\cdot,\varepsilon)\|_{\rho-\delta} \leq r^* \exp(-(m/|\varepsilon|)^{1/\gamma}\delta)$ $\forall\, \delta \in\,]0,\rho]$.*

Furthermore, the quasi-periodic change of variables that performs this transformation is also an element of $\mathcal{Q}_d(\rho, \omega)$. Finally, a general explicit computation of ε^*, a^*, r^*, and m is possible:

$$\varepsilon^* = \min\left(\varepsilon_0, \frac{\alpha}{eq\beta(3d-1)}\right), \qquad a^* = \frac{eq\beta^2}{e-1}, \qquad r^* = ea^*, \qquad m = \frac{c}{10eq\beta}$$

where $e = \exp(1)$, $\alpha = \min_{j\neq\ell}(|\lambda_j - \lambda_\ell|)$, and β is the condition number of a regular matrix S such that $S^{-1}AS$ is diagonal, that is, $\beta = C(S) = |S^{-1}|_\infty |S|_\infty$.

Remark 3.1. For fixed values of $\lambda_1, \ldots, \lambda_d$ and γ, hypothesis 3 is not satisfied for any $c > 0$ only for a set of values of ω of zero measure if $\gamma > r - 1$.

Remark 3.2. In case the eigenvalues of the perturbed matrices move on balls of radius $\mathcal{O}(\varepsilon^p)$ (that is, if the nondegeneracy hypothesis needed in [8] or [9] is not satisfied), it is not difficult to show that the bound of the exponential can be improved: $\|R^*(\cdot, \varepsilon)\|_{\rho-\delta} \leq r^* \exp(-(m/|\varepsilon|)^{p/\gamma}\delta)$. The proof is very similar but uses $M(\varepsilon) = (m/|\varepsilon|)^{p/\gamma}$ instead of $(m/|\varepsilon|)^{1/\gamma}$.

Remark 3.2 seems to show that this nondegeneracy hypothesis is not necessary, and it is only used for technical reasons. In fact, the results seem to be better when this hypothesis is not satisfied.

Remark 3.3. If the unperturbed matrix A has multiple eigenvalues (that is, if hypothesis 1 is not satisfied), the theorem is still true, but the exponent of ε in the exponential of the remainder is slightly worse. This happens because the (small) divisors are now raised to a power that increases with the multiplicity of the eigenvalues. The proof is not included since it does not introduce new ideas and the technical details are rather tedious.

Remark 3.4. The values of ε^*, a^*, r^*, and m given in the theorem are rather pessimistic. In the proof, we have used simple (but rough) bounds instead of cumbersome but more accurate ones. If one is interested in realistic bounds for a given problem, the best thing to do is to rewrite the proof for that particular case. We have done this in section 6 where, with the help of a computer program, we have applied some steps of the method to an example. This allows us to obtain not only better bounds but also (numerically) the reduced matrix as well as the corresponding change of variables.

4. Lemmas.

We will use some lemmas to simplify the proof of Theorem 3.1.

4.1. Basic lemmas.

LEMMA 4.1. *Let $Q(t) = \sum_{k\in\mathbb{Z}^r} Q_k e^{i(k,\omega)t}$ be an element of $\mathcal{Q}_d(\rho, \omega)$ and $M > 0$. Let us define $\overline{Q} = Q_0$, $\widetilde{Q}(t) = Q(t) - Q_0$,*

$$Q_{\geq M}(t) = \sum_{\substack{k\in\mathbb{Z}^r \\ |k|\geq M}} Q_k e^{i(k,\omega)t},$$

and $\widetilde{Q}_{<M} = \widetilde{Q} - Q_{\geq M}$. Then we have the bounds

1. $|\overline{Q}|_\infty$, $\|\widetilde{Q}\|_\rho$, $\|\widetilde{Q}_{<M}\|_\rho \leq \|Q\|_\rho$ *and*
2. $\|Q_{\geq M}\|_{\rho-\delta} \leq \|Q\|_\rho e^{-M\delta}$ $\forall\, \delta \in\,]0, \rho]$.

Proof. The proof follows immediately. □

The next lemma is used to control the variation of the eigenvalues of a perturbed diagonal matrix.

LEMMA 4.2. *Let D be a $d\times d$ diagonal matrix with different eigenvalues $\lambda_1, \ldots, \lambda_d$ and $\alpha = \min_{j\neq\ell}(|\lambda_j - \lambda_\ell|)$. Then if A verifies $|A - D|_\infty \leq b \leq \alpha/(3d-1)$, the following conditions hold:*

1. *A has different eigenvalues $\mu_1, \dots, \mu_d$ and $|\lambda_j - \mu_j| \le b$ if $j = 1, \dots, d$.*

2. *There exists a regular matrix S such that $S^{-1}AS = D^* = \operatorname{diag}(\mu_1, \dots, \mu_d)$ satisfying $C(S) \le 2$.*

Proof. The proof is contained in [8]. □

LEMMA 4.3. *Let $(q_n)_n$, $(a_n)_n$, and $(r_n)_n$ be sequences defined by*

$$q_{n+1} = q_n^2, \qquad a_{n+1} = a_n + q_{n+1}, \qquad r_{n+1} = \frac{2+q_n}{2-q_n} r_n + q_{n+1}$$

with initial values $q_0 = a_0 = r_0 = e^{-1}$. Then $(q_n)_n$ is decreasing to zero and $(a_n)_n$ and $(r_n)_n$ are increasing and convergent to some values a_∞ and r_∞, respectively, with $a_\infty < 1/(e-1)$ and $r_\infty < e/(e-1)$.

Proof. It is immediate that q_n goes to zero quadratically, and this implies that a_n is convergent to the value a_∞:

$$a_\infty = \sum_{j=0}^{\infty} q_j < \sum_{j=1}^{\infty} e^{-j} = \frac{1}{e-1}.$$

Then

$$r_n \le p \left(r_0 + \sum_{j=1}^{n} q_j \right) \le p a_\infty,$$

where $p = \prod_{j=0}^{\infty} (2+q_j)/(2-q_j)$. This product is convergent. In fact,

$$\ln p = \sum_{j=0}^{\infty} \left[\ln\left(1 + \frac{q_j}{2}\right) - \ln\left(1 - \frac{q_j}{2}\right) \right] \le \frac{3}{2} a_\infty \le \frac{3}{2(e-1)} < 1,$$

and so $p < e$, where we have used the fact that $\ln(1+x) \le x$ and $-\ln(1-x) \le 2x$ for $x \in (0, 1/2)$. □

4.2. The inductive lemma. The next lemma is used to perform a step of the inductive procedure.

Before stating the result, let us introduce some notation. Let D and α be as in Lemma 4.2 and let ε^*, q^*, L, and $M(\varepsilon)$ be positive constants. We consider the equation at the step n of the iterative process:

$$\dot{x}_n = (A_n(\varepsilon) + \varepsilon Q_n(t,\varepsilon) + \varepsilon R_n(t,\varepsilon)) x_n, \quad |\varepsilon| \le \varepsilon^*, \tag{8}$$

where $Q_n(\cdot,\varepsilon)$, $R_n(\cdot,\varepsilon) \in \mathcal{Q}_d(\rho,\omega)$ and $\overline{Q}_n(\varepsilon) = Q_n(\cdot,\varepsilon)_{\ge M(\varepsilon)} = 0$. We assume that for some a_n, q_n, $r_n \ge 0$ and $|\varepsilon| < \varepsilon^*$, the following bounds hold:

$$|A_n(\varepsilon) - D| \le q^* a_n |\varepsilon|, \qquad \|Q_n(\cdot,\varepsilon)\|_\rho \le q^* q_n, \qquad \|R_n(\cdot,\varepsilon)\|_{\rho-\delta} \le q^* r_n e^{-M(\varepsilon)\delta},$$

where δ is such that $0 < \delta \le \rho$. (The constant q^* has been introduced to simplify the proof of the theorem later on.) We want to see if it is possible to apply a step of the iterative process to equation (8) to obtain

$$\dot{x}_{n+1} = (A_{n+1}(\varepsilon) + \varepsilon Q_{n+1}(t,\varepsilon) + \varepsilon R_{n+1}(t,\varepsilon)) x_{n+1}, \quad |\varepsilon| \le \varepsilon^*, \tag{9}$$

such that $Q_{n+1}(\cdot,\varepsilon)$, $R_{n+1}(\cdot,\varepsilon) \in \mathcal{Q}_d(\rho,\omega)$ and $\overline{Q}_{n+1}(\varepsilon) = Q_{n+1}(\cdot,\varepsilon)_{\ge M(\varepsilon)} = 0$. We also want to relate the bounds a_{n+1}, q_{n+1}, and r_{n+1} of the terms of this equation with the corresponding bounds of equation (8).

LEMMA 4.4. *Let $\lambda_1^{(n)}(\varepsilon), \dots, \lambda_d^{(n)}(\varepsilon)$ be the eigenvalues of $A_n(\varepsilon)$. Under the previous notation, if*

1. *$L \geq 8q^*$, $\varepsilon^* \leq \alpha/q^*(3d-1)$,*
2. *$a_n \leq 1$, $q_n \leq e^{-1}$, and*
3. *the condition*

$$|\lambda_j^{(n)}(\varepsilon) - \lambda_\ell^{(n)}(\varepsilon) + i(k,\omega)| \geq L|\varepsilon|, \quad |\varepsilon| \leq \varepsilon^*,$$

is satisfied for all j and ℓ and for all $k \in \mathbb{Z}^r$ such that $0 < |k| < M(\varepsilon)$, then equation (8) can be transformed into (9) and

$$q_{n+1} = q_n^2, \qquad a_{n+1} = a_n + q_{n+1}, \qquad r_{n+1} = \frac{2+q_n}{2-q_n} r_n + q_{n+1}\,.$$

The quasi-periodic change of variables that performs this transformation is

$$x_n = (I + \varepsilon P_n(t,\varepsilon))x_{n+1}, \tag{10}$$

where $P_n(\cdot,\varepsilon)$ is the (only) solution of

$$\dot{P}_n = A_n(\varepsilon)P_n - P_n A_n(\varepsilon) + Q_n(t,\varepsilon), \quad \overline{P}_n = 0, \tag{11}$$

that belongs to $\mathcal{Q}_d(\rho,\omega)$. Moreover, $\|\varepsilon P_n(\cdot,\varepsilon)\|_\rho \leq q_n/2 < 1/2$.

Remark 4.1. A_n, Q_n, R_n, P_n, M, and $\lambda_j^{(n)}$ depend on ε but, for simplicity, we will not write this explicitly.

Proof of Lemma 4.4. Let us begin by studying the solutions of (11). Let S_n be the matrix found in Lemma 4.2 with $S_n^{-1}A_nS_n = D_n = \operatorname{diag}(\lambda_1^{(n)},\ldots,\lambda_d^{(n)})$ and $C(S_n) \leq 2$. This lemma can be applied because

$$|A_n - D|_\infty \leq q^* a_n|\varepsilon| \leq q^*\varepsilon^* \leq \frac{\alpha}{3d-1} \quad \forall\, |\varepsilon| \leq \varepsilon^*.$$

Making the change of variables $P_n = S_nX_nS_n^{-1}$ and defining $Y_n = S_n^{-1}Q_nS_n$, equation (11) becomes

$$\dot{X}_n = D_nX_n - X_nD_n + Y_n, \quad \overline{Y}_n = 0.$$

Since D_n is a diagonal matrix, we can handle this equation as d^2 unidimensional equations, which can be easily solved by expanding in Fourier series. If $X_n = (x_{\ell j,n})$ and $Y_n = (y_{\ell j,n})$ with

$$x_{\ell j,n}(t) = \sum_{\substack{k\in\mathbb{Z}^r\\0<|k|<M}} x_{\ell j,n}^k e^{i(k,\omega)t}, \qquad y_{\ell j,n}(t) = \sum_{\substack{k\in\mathbb{Z}^r\\0<|k|<M}} y_{\ell j,n}^k e^{i(k,\omega)t},$$

the coefficients must be

$$x_{\ell j,n}^k = \frac{y_{\ell j,n}^k}{\lambda_j^{(n)} - \lambda_\ell^{(n)} + i(k,\omega)},$$

and by hypothesis 3, they can be bounded by $|x_{\ell j,n}^k| \leq (L|\varepsilon|)^{-1}|y_{\ell j,n}^k|$, which implies

$$\begin{aligned}\|P_n\|_\rho &\leq C(S_n)\|X_n\|_\rho \leq C(S_n)(L|\varepsilon|)^{-1}\|Y_n\|_\rho \leq C(S_n)^2(L|\varepsilon|)^{-1}\|Q_n\|_\rho\\ &\leq 4(L|\varepsilon|)^{-1}q^*q_n \leq |\varepsilon|^{-1}\frac{q_n}{2}.\end{aligned}$$

Hence $\|\varepsilon P_n\|_\rho \leq q_n/2 < 1/2$. Thus $I + \varepsilon P_n$ is invertible and

$$\|(I + \varepsilon P_n)^{-1}\|_\rho \leq \frac{1}{1 - \|\varepsilon P_n\|_\rho} < 2.$$

Now applying the change of (10) to (8) and defining $Q_n^* = \varepsilon(I+\varepsilon P_n)^{-1} Q_n P_n$, $A_{n+1} = A_n + \varepsilon \overline{Q_n^*}$, $Q_{n+1} = (\widetilde{Q_n^*})_{<M}$, and $R_{n+1} = (I + \varepsilon P_n)^{-1} R_n (I + \varepsilon P_n) + (Q_n^*)_{\geq M}$, it is easy to derive equation (9). Finally, we use Lemma 4.1 to bound the terms of this equation:

$$\begin{aligned}
\|Q_n^*\|_\rho &\leq \|(I + \varepsilon P_n)^{-1}\|_\rho \|Q_n\|_\rho \|\varepsilon P_n\|_\rho \leq \|Q_n\|_\rho q_n \leq q^* q_n^2 = q^* q_{n+1},\\
\|Q_{n+1}\|_\rho &\leq \|Q_n^*\|_\rho \leq q^* q_{n+1},\\
|A_{n+1} - D|_\infty &\leq |A_n - D|_\infty + |\varepsilon \overline{Q_n^*}|_\infty \leq q^*(a_n + q_{n+1})|\varepsilon| = q^* a_{n+1}|\varepsilon|,\\
\|R_{n+1}\|_{\rho-\delta} &\leq \frac{1 + \|\varepsilon P_n\|_\rho}{1 - \|\varepsilon P_n\|_\rho} \|R_n\|_{\rho-\delta} + \|(Q_n^*)_{\geq M}\|_{\rho-\delta}\\
&\leq \left(\frac{1 + q_n/2}{1 - q_n/2} r_n + q_{n+1} \right) q^* e^{-M\delta} = q^* r_{n+1} e^{-M\delta} \quad \forall\, \delta \in \,]0, \rho],
\end{aligned}$$

and the proof is complete. □

5. Proof of Theorem 3.1. Let S be a regular matrix such that $S^{-1}AS = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. We define ε^*, α, β, and m as in Theorem 3.1. We also define $q^* = e\beta q$, $M = M(\varepsilon) = (m/|\varepsilon|)^{1/\gamma}$, and $L = 8q^*$.

The (constant) change $x = Sx_0$ transforms the initial equation into

$$\dot{x}_0 = (D + \varepsilon Q^*(t, \varepsilon)) x_0, \tag{12}$$

where $Q^* = S^{-1}QS$ and so $\|Q^*\|_\rho \leq e^{-1} q^*$ for $|\varepsilon| \leq \varepsilon^*$. We split equation (12) as follows:

$$\dot{x} = (A_0 + \varepsilon Q_0(t) + \varepsilon R_0(t)) x_0,$$

where $A_0 = D + \varepsilon \overline{Q^*}$, $Q_0 = \widetilde{Q^*}_{<M}$, and $R_0 = Q^*_{\geq M}$. Using Lemma 4.1, it is easy to see that

$$|A_0 - D|_\infty \leq q^* a_0 |\varepsilon|, \qquad \|Q_0\|_\rho \leq q^* q_0, \qquad \|R_0\|_{\rho-\delta} \leq q^* r_0 e^{-M\delta}$$

$\forall\, \delta \in \,]0, \rho]$ and $|\varepsilon| \leq \varepsilon^*$ if $a_0 = q_0 = r_0 = e^{-1}$.

We will show that in all of the steps, the hypotheses of Lemma 4.4 are satisfied. Since hypothesis 1 and 2 are easy to check, we focus on hypothesis 3.

Now since $a_n \leq 1$, $|\varepsilon| \leq \varepsilon^*$, and $|A_n - D|_\infty \leq q^*|\varepsilon| \leq \alpha/(3d-1)$, Lemma 4.2 gives that

$$|\alpha_{j\ell}^{(n)} - \alpha_{j\ell}| < 2q^*|\varepsilon| \quad \forall\, j,\, \ell,\ |\varepsilon| \leq \varepsilon^*,$$

where $\alpha_{j\ell} = \lambda_j - \lambda_\ell$ and $\alpha_{j\ell}^{(n)} = \lambda_j^{(n)} - \lambda_\ell^{(n)}$, where $\lambda_1^{(n)}, \ldots, \lambda_d^{(n)}$ are the eigenvalues of $A_n(\varepsilon)$.

Using hypothesis 3 of Theorem 3.1, we obtain that if $k \in \mathbb{Z}^r$ and $0 < |k| < M(\varepsilon)$,

$$\begin{aligned}
|\alpha_{j\ell}^{(n)} + i(k, \omega)| &\geq |\alpha_{j\ell} + i(k, \omega)| - |\alpha_{j\ell}^{(n)} - \alpha_{j\ell}| > \frac{c}{|k|^\gamma} - 2q^*|\varepsilon|\\
&> \left(\frac{c}{m} - 2q^* \right) |\varepsilon| = L|\varepsilon|,
\end{aligned}$$

and hypothesis 3 of Lemma 4.4 is verified.

Consequently, the iterative process can be carried out and Lemma 4.3 ensures the convergence of the process. The composition of all of the changes $I + \varepsilon P_n$ is convergent because $\|I + \varepsilon P_n\|_\rho \leq 1 + q_n/2$. Then the final equation is

$$\dot{x}_\infty = (A_\infty(\varepsilon) + \varepsilon R_\infty(t, \varepsilon))x_\infty, \quad |\varepsilon| \leq \varepsilon^*, \tag{13}$$

where $|A_\infty(\varepsilon) - D|_\infty \leq q^* a_\infty |\varepsilon| \leq (e\beta/(e-1))q|\varepsilon|$ and

$$\|R_\infty(\cdot, \varepsilon)\|_{\rho-\delta} \leq q^* r_\infty e^{-M(\varepsilon)\delta} \leq \frac{e^2\beta}{e-1} q \exp\left\{-\left(\frac{m}{|\varepsilon|}\right)^{1/\gamma} \delta\right\} \quad \forall\, \delta \in]0, \rho].$$

To complete the proof, the change $x_\infty = S^{-1}y$ transforms equation (13) into equation (7) with the bounds that we were looking for.

6. An example. The results of this paper can be applied in many ways according to the kind of problem we are interested in. Let us illustrate this with the help of an example.

Let us consider the equation

$$\ddot{x} + (1 + \varepsilon q(t))x = 0, \tag{14}$$

where $q(t) = \cos(\omega_1 t) + \cos(\omega_2 t)$ with $\omega_1 = \sqrt{2}$ and $\omega_2 = \sqrt{3}$. Defining y as $\dot{x}$, we can rewrite (14) as

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \left[\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \varepsilon \begin{pmatrix} 0 & 0 \\ -q(t) & 0 \end{pmatrix}\right] \begin{pmatrix} x \\ y \end{pmatrix}. \tag{15}$$

Since $\lambda_{1,2} = \pm i$, the diophantine condition (6) is satisfied for $\gamma = 1$ (because the frequencies are quadratic irrationals). The value of c will be discussed later. For the sake of simplicity, let us take $\rho = 2$ and $\delta = 1$. This implies that $q = \|Q\|_\rho = 2e^2$. It is not difficult to derive $\beta = 2$ and, finally, $\varepsilon^* = 4.9787\ldots \times 10^{-3}$ and $r^* = 2.5419\ldots \times 10^2$.

The value of c might be calculated for all $k = (k_1, k_2)$, but better (larger) values can be used since we need to consider $|k|$ only up to a finite order. For instance, an easy computation shows that for $|k| \leq 125$, c is 0.149. If $|k| = 126$, then c must be at most 0.013 due to the quasi resonance produced by $k = (70, -56)$. In the range $126 \leq |k| \leq 10^5$, there are no more relevant resonances, so the value $c = 0.013$ suffices.

To begin our discussion, let us suppose that the value of ε in (15) is $\varepsilon = 2 \times 10^{-6}$. If we take $c = 0.149$, we obtain that $m = 1.8545\ldots \times 10^{-4}$ and $M = 93$. (Recall that the process cancels frequencies such that $|k| < M(\varepsilon)$.) If the value of M had been larger than 125, we would have used the value $c = 0.013$ instead. Therefore, we can reduce the system to constant coefficients with a remainder R^* such that $\|R^*\|_{\rho-1} < 10^{-37}$.

If the given value of ε is smaller—for instance, $\varepsilon = 10^{-7}$—the computed value of M if $c = 0.149$ is 1855, so $c = 0.013$ must be used. This produces $M = 162$ and $\|R^*\|_{\rho-1} < 10^{-67}$. A value of $\varepsilon = 5 \times 10^{-8}$ implies that $M = 324$ and $\|R^*\|_{\rho-1} < 10^{-138}$. The computation of the reduced matrix as well as the quasi-periodic change of variables will be discussed below.

Another interesting problem is the study of reducibility for a value of ε larger than the ε^* given above. Let us continue working with the same equation but with $\varepsilon = 0.1$ as our example.

To increase the value of ε^*, one may try to rewrite the proof using optimal bounds at each step. This has not been done here in order to get an easy, clean, and short proof. Instead of doing this, we think that it is much better to rewrite the proof for our example using no bounds but exact values. This will produce the best results for this problem.

For that purpose, we have implemented the algorithm used in the proof of the theorem as a C program for a (given) fixed value of ε. The program computes and performs a finite number of the changes of variables used to prove the theorem. As a result, the reduced system (including the remainder) as well as the final change of variables are written.

To simplify and make the program more efficient, all of the coefficients have been stored as double-precision variables. During all of the operations, all of the coefficients less than 10^{-20} have been dropped in order to control the size of the Fourier series that appears during the process. Of course, this introduces some (small) numerical error in the results.[1]

After four changes of variables, (15) is transformed into

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \left[\begin{pmatrix} 0.0 & b_{12} \\ b_{21} & 0.0 \end{pmatrix} + R(t) \right] \begin{pmatrix} x \\ y \end{pmatrix}, \tag{16}$$

where $b_{12} = 1.000000366251255$ and $b_{21} = -0.992421151834871$. The remainder R is very small: the largest coefficient it contains is less than 10^{-16}. Note that the accuracy (relative error) of this remainder is very poor due to the use of double-precision arithmetic (15–16 digits) for the coefficients. During the computations, M has not been given a value. Instead, we have tried to cancel all the frequencies with amplitude larger than 10^{-16}. (It turns out from the computations that all of these frequencies satisfy $|k| \leq 20$.) It is also possible to obtain a better accuracy in the result, using a multiple-precision arithmetic.

Finally, to check the software, we have tabulated a solution of (16) for a timespan of 10 time units. We have transformed this table by means of the (quasi-periodic) change of variables given by the program. Then we have taken the first point of the transformed table as initial condition of (15) to produce (by means of numerical integration) a new table. The differences between these two tables are less than 10^{-13}, as expected.

Therefore, for practical purposes, this is an "effective" Floquet theorem in the sense that it allows to compute the reduced matrix as well as the change of variables with the usual accuracy used in numerical computations.

Acknowledgments. The authors want to thank C. Bonet and C. Simó for fruitful discussions and remarks.

REFERENCES

[1] V. I. Arnol'd, *Proof of a theorem of A. N. Kolmogorov on the invariance of quasi-periodic motions under small perturbations of the Hamiltonian*, Russian Math. Surveys, 18 (1963), pp. 9–36.

[2] N. N. Bogoljubov, Ju. A. Mitropoliski, and A. M. Samoilenko, *Methods of Accelerated Convergence in Nonlinear Mechanics*, Springer-Verlag, New York, 1976.

[3] L. Chierchia, *Absolutely continuous spectra of quasiperiodic Schrödinger operators*, J. Math. Phys., 28 (1987), pp. 2891–2898.

[1] If one wants to control that error, it is possible to use intervalar arithmetic for the coefficients and carry a bound of the remainder for each Fourier series.

[4] E. I. Dinaburg and J. G. Sinai, *The one-dimensional Schrödinger equation with quasiperiodic potential*, Funct. Anal. Appl., 9 (1975), pp. 8–21.

[5] L. H. Eliasson, *Floquet solutions for the* 1*-dimensional quasi-periodic Schrödinger equation*, Comm. Math. Phys., 146 (1992), pp. 447–482.

[6] A. M. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Math. 377, Springer-Verlag, Berlin, 1974.

[7] R. A. Johnson and G. R. Sell, *Smoothness of spectral subbundles and reducibility of quasi-periodic linear differential systems*, J. Differential Equations, 41 (1981), pp. 262–288.

[8] A. Jorba and C. Simó, *On the reducibility of linear differential equations with quasiperiodic coefficients*, J. Differential Equations, 98 (1992), pp. 111–124.

[9] A. Jorba and C. Simó, *On quasi-periodic perturbations of elliptic equilibrium points*, SIAM J. Math. Anal., 27 (1996), pp. 1704–1737.

[10] J. Moser and J. Pöschel, *On the stationary Schrödinger equation with a quasiperiodic potential*, Phys. A, 124 (1984), pp. 535–542.

[11] J. Moser and J. Pöschel, *An extension of a result by Dinaburg and Sinai on quasi-periodic potentials*, Comment. Math. Helv., 59 (1984), pp. 39–85.

[12] A. I. Neishtadt, *The separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech., 48 (1984), pp. 133–139.

[13] H. Rüssmann, *On the one-dimensional Schrödinger equation with a quasi-periodic potential*, Ann. New York Acad. Sci., 357 (1980), pp. 90–107.

[14] C. Simó, *Averaging under fast quasiperiodic forcing*, in Hamiltonian Mechanics: Integrability and Chaotic Behaviour, NATO ASI Series B: Physics, Plenum Press, New York, 1994, pp. 13–34.

[15] D. Treshchev, *An estimate of irremovable nonconstant terms in the reducibility problem*, Amer. Math. Soc. Transl. Ser. 2, 168 (1995), pp. 91–128.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 189–212, January 1997

010

FUNCTION NORMS AND FRACTAL DIMENSION*

CLAUDE TRICOT†

Abstract. Using functional norms $L^\alpha(f)$, we introduce a two-parameter norm family $\mathcal{L}^{(\alpha,\beta)}(f)$ by performing sections on the definition domain of f. These norms are used on the difference function $f(x)-f(y)$ to obtain the operators $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ which measure the irregularity of f. The order of growth of $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ at 0 determines an irregularity index $\Delta^{(\alpha,\beta)}(f)$. In particular, $\Delta^{(\infty,1)}(f)$ is the fractal dimension of the graph of f. We investigate the value of $\Delta^{(\alpha,\beta)}(f)$ for the series $f(x) = \sum_{n=0}^{\infty} 2^{-nH} g(2^n x + \phi_n)$, where $0 < H < 1$, (ϕ_n) is a real-number sequence, and g is a continuous periodic function of period 1.

Key words. norm, fractal dimension, functional norm, oscillation, irregular function

AMS subject classifications. 28A75, 28A80, 46E30, 26A16, 47A30

PII. S0036141094278791

1. Introduction.

In signal analysis, typical data sets can be represented by irregular functions. There are many ways to give a mathematical meaning to the word "irregular." For example, we can say that a function is irregular if it is nowhere differentiable. However, this criterion does not allow us to compare the irregularities of two functions. We can define the *irregularity degree* of a function f defined in a domain D of R^N by evaluating the fractal dimension of its graph G_f. For this evaluation, the procedure described in [11] has been used in many further studies of profiles or surfaces [4, 10]. It consists of calculating for every $x \in D$ and every $\tau > 0$ the *oscillation* of f over the ball $B_\tau(x)$, that is, the difference betwen the supremum and minimum of $f(y)$ for $\|x - y\| \leq \tau$. The arithmetic mean of these oscillations over D is a function of τ, called the *τ-variation* V_τ of f. When τ tends to 0, the order of growth of V_τ is directly related to the graph dimension (in the sense of Minkowski and Bouligand), which we denote by $\mathrm{Dim}(G_f)$.

Many functions other than V_τ may be used to characterize the irregularity of f. As a general rule, we first perform a *local analysis* of the irregularity on the ball $B_\tau(x)$ and then use an averaging process over D to obtain some *global* measure. A number of researchers prefer to discard the oscillation measurements since the extremal values of a signal may be too sensitive to errors in the data acquisition. They may instead choose an average of the differences $|f(x) - f(y)|$ in the neighborhood of x. If we choose, for example, to perform a quadratic mean both locally (over $B_\tau(x)$) and globally (over D), we obtain the *standard deviation* of $|f(x) - f(y)|$ over the whole definition domain, that is, $\mathcal{D}_\tau = \{(x, y) \in R^{2N} \,/\, x \in \mathrm{D},\, \|x-y\| \leq \tau\}$. In this analysis, the behavior of the function

$$\sqrt{\frac{1}{(2\tau)^N \,\mathrm{Vol}_N(\mathrm{D})} \int_{\mathrm{D}} \left(\int_{B_\tau(x)} |f(x) - f(y)|^2 \, dy \right) dx}$$

at $\tau = 0$ is used to characterize the irregularity of f.

* Received by the editors December 16, 1994; accepted for publication (in revised form) October 13, 1995.

http://www.siam.org/journals/sima/28-1/27879.html

† Département de Mathématiques Appliquées, Ecole Polytechnique de Montréal, C.P. 6079, Succursale Centre-Ville, Montréal, PQ H3C 3A7, Canada (claude@graf.polymtl.ca).

One of our aims is to gather various such irregularity functions into the same family and to find their relationships. To achieve this goal, the seminorms L^α, $\alpha \geq 1$, are introduced in section 2. By using cross-sections of the definition domain D of f, we show in section 3 how to construct a doubly indexed family $\mathcal{L}^{(\alpha,\beta)}$ of seminorms for $\alpha \in [1, +\infty]$ and $\beta \in [1, +\infty]$. Then in section 4, we define our general irregularity functions $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$. For example,

$$\mathrm{S}_\tau^{(\alpha,\alpha)}(f) = \left(\frac{1}{\mathrm{Vol}_{2N}(\mathcal{D}_\tau)} \int_{\mathcal{D}_\tau} |f(x) - f(y)|^\alpha \, dy \, dx \right)^{\frac{1}{\alpha}}.$$

The case where $\alpha = \beta = 2$ corresponds to the standard deviation. Another example is

$$\mathrm{S}_\tau^{(\infty,1)}(f) = \frac{1}{\mathrm{Vol}_N(\mathrm{D})} \int_{\mathrm{D}} \left(\sup_{y \in B_\tau(x)} |f(x) - f(y)| \right) dx,$$

a function equivalent to the variation of f. Also,

$$\mathrm{S}_\tau^{(\infty,\infty)}(f) = \sup_{(x,y) \in \mathcal{D}_\tau} |f(x) - f(y)|$$

is the maximum oscillation of f over D.

We study the main properties of the operators $\mathrm{S}_\tau^{(\alpha,\beta)}$ and show that they are continuous and increasing with respect to each of the variables α and β. For each of them, we introduce in section 5 an *irregularity index* as follows:

$$\Delta^{(\alpha,\beta)}(f) = \limsup_{\tau \to 0} \left(N + 1 - \frac{\log \mathrm{S}_\tau^{(\alpha,\beta)}(f)}{\log \tau} \right).$$

The special case $\Delta^{(\infty,1)}(f)$ gives the dimension of the graph of f, but all of these indices are interesting in themselves for the characterization of irregularities. For differentiable functions, they all have the same value N. For many *fractal* functions, they still take a common value larger than N. However, they are not identical in general since for some f, the irregularity functions $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ are not equivalent near 0. We show that $\Delta^{(\alpha,\beta)}(f)$ as a function of α and β is increasing on $[1, +\infty] \times [1, +\infty]$ and continuous on $[1, +\infty) \times [1, +\infty)$. Finally, in section 6, we study a well-known family of nowhere-differentiable functions defined by the series

$$f(x) = \sum_{n=0}^{+\infty} 2^{-nH} g(2^n x + \phi_n),$$

where $0 < H < 1$, $\phi_n \in R$, and g is continuous and periodic with period 1. To simplify the arguments, we consider only the functions g that verify the extra property

$$g(x) + g\left(x + \frac{1}{2}\right) = \text{constant}.$$

Particular cases of such functions f are the Weierstrass and the Knopp–Takagi functions. We give the conditions on g so that our irregularity indices $\Delta^{(\alpha,\beta)}(f)$ all take the same value $2 - H$ (for the same subject in a slightly different setting, see, e.g., [8, 1]).

We also show that for other fuctions g, the indices may take a common value different from $2-H$ or even take values that depend on the pair (α, β).

One application of this paper consists of providing approximations of the dimension $\mathrm{Dim}(G_f)$ using other indices that may be more robust numerically. However, our aim is really to exhibit a full range of irregularity indices and to show their relationships in order to give the necessary theoretical support to their experimental utilization.

2. $(\boldsymbol{\alpha})$ seminorms.

2.1. Definitions. Let D be a measurable set in R^N, $N \geq 2$, and $\mathrm{Vol}_N(\mathrm{D})$ be its N-dimensional Lebesgue measure. We assume that $0 < \mathrm{Vol}_N(\mathrm{D}) < +\infty$. Let $\mathcal{M}_b(\mathrm{D})$ be the set of all bounded, measurable functions $f : \mathrm{D} \longrightarrow R$. For each $f \in \mathcal{M}_b(\mathrm{D})$ and for each $\alpha > 0$, let us define the following multiple integral:

$$L^\alpha(f;\mathrm{D}) = \left(\frac{1}{\mathrm{Vol}_N(\mathrm{D})}\int_{\mathrm{D}} |f(x)|^\alpha \, dx\right)^{1/\alpha}, \tag{1}$$

where $x = (x_1, \ldots, x_N)$. If $\alpha \geq 1$, L^α is a *seminorm* on $\mathcal{M}_b(\mathrm{D})$. The extra term $1/\mathrm{Vol}_N(\mathrm{D})$ constitutes a slight change with respect to the classical seminorm definition. Its introduction allows us to consider L^α as an *average* and implies that $L^\alpha(f;\mathrm{D})$ is increasing with respect to α (see Theorem 2.1 below), a crucial property in this paper.

We complete (1) with the standard definition of the seminorm L^∞, the *essential supremum* of f over its definition domain:

$$L^\infty(f;\mathrm{D}) = \operatorname{ess\,sup}_{x\in\mathrm{D}}|f(x)|. \tag{2}$$

For fixed D and f, $L^\alpha(f;\mathrm{D})$ may be considered as a function of the variable α defined on $]0,+\infty]$, the set of positive, nonzero, real numbers completed with $+\infty$. This will allow us in section 2.3 to consider the continuity of L^α at $+\infty$.

2.2. Seminorms on the cross-sections. Let $N \geq 2$. We may also define the (α) seminorms of f when restricted to a cross-section of D by a hyperplane.

Given $k < N$ and $t = (t_1, \ldots, t_k)$, a point of the projection of D on R^k, we denote by

$$\mathrm{D}^{(t)} = \mathrm{D} \cap \{\{t\} \times R^{N-k}\} \tag{3}$$

the cross-section of D by the hyperplane of equations $x_1 = t_1, \ldots, x_k = t_k$. Let us consider $\mathrm{D}^{(t)}$ as a subset of R^{N-k} and assume that $0 < \mathrm{Vol}_{N-k}(\mathrm{D}^{(t)}) < +\infty$. Then for all $\alpha > 0$, we define

$$L^\alpha(f;\mathrm{D}^{(t)}) = \left(\frac{1}{\mathrm{Vol}_{N-k}(\mathrm{D}^{(t)})}\int_{\mathrm{D}^{(t)}} |f(x)|^\alpha \, dx_{k+1}\cdots dx_N\right)^{1/\alpha}, \tag{4}$$

where $x = (t_1, \ldots, t_k, x_{k+1}, \ldots, x_N)$. Also,

$$L^\infty(f;\mathrm{D}^{(t)}) = \operatorname{ess\,sup}_{x\in\mathrm{D}^{(t)}}|f(x)|. \tag{5}$$

2.3. Properties. Let us enumerate some fundamental properties of the operator $L^\alpha(f;\mathrm{D})$. They are verified by $L^\alpha(f;\mathrm{D}^{(t)})$ as well. The first three characterize a seminorm.

Theorem 2.1. *The operators L^α have the following properties:*

1. $L^\alpha(f;\mathrm{D}) = 0 \iff f(x) = 0$ *almost everywhere on* D.
2. *For every a,* $L^\alpha(a\,f;\mathrm{D}) = |a|\,L^\alpha(f;\mathrm{D})$.
3. *If f_1, $f_2 \in \mathcal{M}_b(\mathrm{D})$ and $\alpha \in [1,+\infty]$, then*

$$L^\alpha(f_1+f_2;\mathrm{D}) \le L^\alpha(f_1;\mathrm{D}) + L^\alpha(f_2;\mathrm{D}). \tag{6}$$

This is the triangular inequality *associated with the Minkowski inequality for finite sums.*

4. *If f_1, $f_2 \in \mathcal{M}_b(\mathrm{D})$ and if $f_1(x) \le f_2(x)$ almost everywhere on* D*, then*

$$L^\alpha(f_1;\mathrm{D}) \le L^\alpha(f_2;\mathrm{D}). \tag{7}$$

5. *Fix $f \in \mathcal{M}_b(\mathrm{D})$. Then $L^\alpha(f;\mathrm{D})$ is a continuous, increasing function of α on $]0,+\infty]$.*
6. *For all α, $\beta \in]0,+\infty]$,*

$$L^\alpha(|f|^\beta;\mathrm{D}) = (L^{\alpha\beta}(f;\mathrm{D}))^\beta. \tag{8}$$

7. *For all $\alpha \in]0,+\infty]$, $\beta \in]0,+\infty]$, and $\beta \le \alpha$,*

$$L^\beta(f;\mathrm{D}) \le L^\alpha(f;\mathrm{D}) \le (L^\beta(f;\mathrm{D}))^{\frac{\beta}{\alpha}}(L^\infty(f;\mathrm{D}))^{1-\frac{\beta}{\alpha}}. \tag{9}$$

The proof of formula (9) uses the inequality

$$\int_\mathrm{D} |f(x)|^\alpha\,dx \le \left(\int_\mathrm{D} |f(x)|^\beta\,dx\right) \operatorname{ess\,sup}_\mathrm{D}|f(x)|^{\alpha-\beta},$$

which is true when $\beta \le \alpha$. The proof of the other properties uses classical arguments and they are left to the reader.

3. $(\boldsymbol{\alpha},\boldsymbol{\beta})$ seminorms.

3.1. Definitions. Let k be an integer, $1 \le k < N$, and D be a measurable subset of R^N with the following two properties:

(i) $0 < \mathrm{Vol}_N(\mathrm{D}) < +\infty$.
(ii) If E is the orthogonal projection of D over R^k, then for all $t \in E$,

$$0 < \mathrm{Vol}_{N-k}(\mathrm{D}^{(t)}) < +\infty.$$

The cross-section $\mathrm{D}^{(t)}$ is defined in section 2.2. We remark that $\mathrm{Vol}_k(E) > 0$. An example of such a set D is any bounded, open set in R^N. Another example will be presented in section 4.1. The two conditions above allow us to calculate the seminorms of f over all cross-sections of D using formula (4).

Given $f \in \mathcal{M}_b(\mathrm{D})$, we may evaluate $L^\alpha(f;\mathrm{D}^{(t)})$ for all $t \in E$. The result is a function of t. We may evaluate the seminorm of this function as well. We finally obtain two-parameter operators as follows:

$$\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D}) = L^\beta(L^\alpha(f;\mathrm{D}^{(t)});E), \tag{10}$$

where $(\alpha,\beta) \in]0,+\infty] \times]0,+\infty]$. As a particular case,

$$\mathcal{L}^{(\infty,\infty)}(f;\mathrm{D}) = L^\infty(f;\mathrm{D}) = \operatorname{ess\,sup}_\mathrm{D}|f(x)|. \tag{11}$$

3.2. Properties. The following is a direct consequence of Theorem 2.1.

THEOREM 3.1. *The operators $\mathcal{L}^{(\alpha,\beta)}$ have the following properties:*

1. $\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D}) = 0 \iff f(x) = 0$ *almost everywhere on* D.
2. *For all a,* $\mathcal{L}^{(\alpha,\beta)}(a\, f;\mathrm{D}) = |a|\,\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D})$.
3. *If f_1, $f_2 \in \mathcal{M}_b(\mathrm{D})$ and $\alpha \in [1,+\infty]$, $\beta \in [1,+\infty]$, then*

$$\mathcal{L}^{(\alpha,\beta)}(f_1+f_2;\mathrm{D}) \le \mathcal{L}^{(\alpha,\beta)}(f_1;\mathrm{D}) + \mathcal{L}^{(\alpha,\beta)}(f_2;\mathrm{D}). \tag{12}$$

These three properties prove that $\mathcal{L}^{(\alpha,\beta)}$ is a seminorm for $(\alpha,\beta) \in [1,+\infty] \times [1,+\infty]$.

4. *If $f \in \mathcal{M}_b(\mathrm{D})$ is fixed, then $\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D})$ is a continuous, increasing function of α and β on $]0,+\infty] \times]0,+\infty]$.*
5. *Let us fix* D *and $f \in \mathcal{M}_b(\mathrm{D})$ and write*

$$\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D}) = \mathcal{L}^{(\alpha,\beta)}, \qquad \mathcal{L}^{(\infty,\infty)}(f;\mathrm{D}) = \|f\|$$

for simplicity. Given real numbers α, β, γ, and δ in $]0,+\infty)$, we seek relationships between $\mathcal{L}^{(\alpha,\beta)}$ and $\mathcal{L}^{(\gamma,\delta)}$. Here are some inequalities:

If $\gamma \le \alpha$,

$$\mathcal{L}^{(\gamma,\beta)} \le \mathcal{L}^{(\alpha,\beta)} \le \left(\mathcal{L}^{(\gamma,\frac{\beta\gamma}{\alpha})}\right)^{\frac{\gamma}{\alpha}} \|f\|^{1-\frac{\gamma}{\alpha}}. \tag{13}$$

If $\delta \le \beta$,

$$\mathcal{L}^{(\alpha,\delta)} \le \mathcal{L}^{(\alpha,\beta)} \le \left(\mathcal{L}^{(\alpha,\delta)}\right)^{\frac{\delta}{\beta}} \|f\|^{1-\frac{\delta}{\beta}}. \tag{14}$$

If $\gamma \le \alpha$ and $\delta \le \beta$,

$$\mathcal{L}^{(\gamma,\delta)} \le \mathcal{L}^{(\alpha,\beta)} \le \left(\mathcal{L}^{(\gamma,\frac{\gamma\delta}{\alpha})}\right)^{\frac{\gamma\delta}{\alpha\beta}} \|f\|^{1-\frac{\gamma\delta}{\alpha\beta}}. \tag{15}$$

Finally, if $\gamma \le \alpha$ and $\beta \le \delta$,

$$\mathcal{L}^{(\gamma,\delta)} \le \left(\mathcal{L}^{(\alpha,\beta)}\right)^{\frac{\beta}{\delta}} \|f\|^{1-\frac{\beta}{\delta}} \le \left(\mathcal{L}^{(\gamma,\frac{\gamma\delta}{\alpha})}\right)^{\frac{\beta\gamma}{\alpha\delta}} \|f\|^{1-\frac{\beta\gamma}{\alpha\delta}}. \tag{16}$$

Inequalities corresponding to the cases where $\alpha \le \gamma$, $\delta \le \beta$ and $\alpha \le \gamma$, $\beta \le \delta$ are deduced from (15) and (16) by a change of variables.

6. *For a fixed $f \in \mathcal{M}_b(\mathrm{D})$, $\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D})$ is continuous on $]0,+\infty] \times]0,+\infty]$.*

Proof. Let us prove (13). The first inequality comes from the monotonicity of $\mathcal{L}^{(\alpha,\beta)}$. The second is obtained by using (9). For all $t \in E$,

$$L^{\alpha}(f;\mathrm{D}^{(t)}) \le (L^{\gamma}(f;\mathrm{D}^{(t)}))^{\frac{\gamma}{\alpha}} (L^{\infty}(f;\mathrm{D}^{(t)}))^{1-\frac{\gamma}{\alpha}}.$$

Also, $L^{\infty}(f;\mathrm{D}^{(t)}) \le \|f\|$. Now using the seminorm L^{β}, we get

$$L^{\beta}(L^{\alpha}(f;\mathrm{D}^{(t)});E) \le L^{\beta}((L^{\gamma}(f;\mathrm{D}^{(t)}))^{\frac{\gamma}{\alpha}};E)\,\|f\|^{1-\frac{\gamma}{\alpha}}.$$

From (8), the right-hand side member of the above equation is simply

$$L^{\frac{\beta\gamma}{\alpha}}(L^{\gamma}(f;\mathrm{D}^{(t)});E)^{\frac{\gamma}{\alpha}}\,\|f\|^{1-\frac{\gamma}{\alpha}}.$$

Hence we have formula (13).

For (14), use (9). For all $g \in \mathcal{M}_b(E)$,

$$L^\delta(g;E) \leq L^\beta(g;E) \leq (L^\delta(g;E))^{\frac{\delta}{\beta}} (L^\infty(g;E))^{1-\frac{\delta}{\beta}}.$$

Then replace $g(t)$ by $L^\alpha(f;\mathrm{D}^{(t)}))$.

Formulas (15) and (16) are obtained from (13) and (14). We may notice that formulas (13)–(16) are all consequences of (9). We could get other formulas by instead using the Hölder inequality for integrals.

For the continuity of $\mathcal{L}^{(\alpha,\beta)}(f;\mathrm{D})$, $\alpha < \infty$, use (13) to get

$$\mathcal{L}^{(\gamma,\beta)} \leq \mathcal{L}^{(\alpha,\beta)} \leq (\mathcal{L}^{(\gamma,\beta)})^{\frac{\gamma}{\alpha}} \|f\|^{1-\frac{\gamma}{\alpha}}. \tag{17}$$

This implies that

$$\lim_{\gamma\to\alpha,\gamma<\alpha} \mathcal{L}^{(\gamma,\beta)} = \mathcal{L}^{(\alpha,\beta)}.$$

This helps us to prove continuity on the left of α and on the right of γ.

Finally, use the Lebesgue convergence theorem to prove continuity at $\alpha = \infty$.

The other elements of the proof are left to the reader. □

4. Norms for the difference function.

4.1. Difference function. Given a measurable subset $\mathrm{D} \subset R^N$ and a nonconstant function $f \in \mathcal{M}_b(\mathrm{D})$, the *difference function* is defined as

$$F(x,y) = f(x) - f(y). \tag{18}$$

Its definition domain is $\mathrm{D} \times \mathrm{D} \subset R^{2N}$, but we are only interested in values of F for x and y close to each other. Only these values can help measure the irregularity of f.

Here are some notations:

If $x = (x_1, \dots, x_N)$ is in R^N, then $\|x\| = \max_{1\leq i\leq N} |x_i|$ is the *norm of the maximum* of x.

τ_0 is a fixed real number > 0 and τ is a parameter in $]0, \tau_0]$.

$B_\tau(x)$ is the ball of center x and radius τ:

$$B_\tau(x) = \{y \,/\, \|x-y\| \leq \tau\}.$$

Its volume is $(2\tau)^N$.

D is a compact subset of R^N such that $\mathrm{Vol}_N(\mathrm{D}) > 0$.

$\mathrm{D}(\tau)$ is the τ-*Minkowski sausage* of D, the compact set

$$\mathrm{D}(\tau) = \cup_{x\in\mathrm{D}} B_\tau(x)$$

of all points in R^N at a distance $\leq \tau$ from D.

$\mathcal{D}_\tau$ is the compact set

$$\mathcal{D}_\tau = \{(x,y) \in R^{2N} \,/\, x \in \mathrm{D},\, \|x-y\| \leq \tau\}. \tag{19}$$

It is included in $\mathrm{D} \times \mathrm{D}(\tau)$. For all $x \in \mathrm{D}$, its x-cross-section is $\mathrm{D}_\tau^{(x)}$ (see equation (3)), identified here with $B_\tau(x)$. We may write

$$\mathcal{D}_\tau = \cup_{x\in\mathrm{D}} \{x\} \times B_\tau(x).$$

This set verifies the two conditions stated in section 3.1. It is therefore possible to calculate the operators $\mathcal{L}^{(\alpha,\beta)}$ of a function defined on $\mathcal{D}_\tau$.

$\mathcal{C}(\mathrm{D})$ is the set of all continuous functions defined on D. Then $\mathcal{C}(\mathrm{D}) \subset \mathcal{M}_b(\mathrm{D})$.

Henceforth, let us assume that the function f is in $\mathcal{C}(\mathrm{D}(\tau_0))$. Therefore, $F \in \mathcal{C}(\mathcal{D}_\tau)$ for all $\tau \leq \tau_0$. The continuity assumption helps us to simplify the vocabulary. For example, the seminorm $\operatorname{ess\,sup}(f)$ is the same as the *norm* $\sup(f)$. The seminorms L^α, $\alpha \geq 1$, and $\mathcal{L}^{(\alpha,\beta)}$, $(\alpha,\beta) \in [1,+\infty] \times [1,+\infty]$, become norms as well.

One method for measuring the *local* irregularity of f consists of evaluating the (α) norm of the difference function $F(x,y)$ over the cross-section $\mathcal{D}_\tau^{(x)}$. The result is a function of x since the local irregularity may vary on D. We get a *global* measure of the irregularity by calculating the (β) norm of this function. The final result is denoted by $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$, or $\mathrm{S}_\tau^{(\alpha,\beta)}$ when there is no ambiguity. For all $(\alpha,\beta) \in]0,+\infty] \times]0,+\infty]$, we have

$$\mathrm{S}_\tau^{(\alpha,\beta)}(f) = \mathcal{L}^{(\alpha,\beta)}(F;\mathcal{D}_\tau) = L^\beta(L^\alpha(F;\mathcal{D}_\tau^{(x)});\mathrm{D}). \tag{20}$$

When $0 < \alpha < +\infty$ and $0 < \beta < +\infty$, this gives

$$\mathrm{S}_\tau^{(\alpha,\beta)}(f) = \left(\frac{1}{(2\tau)^{\frac{N\beta}{\alpha}} \operatorname{Vol}_N(\mathrm{D})} \int_{\mathrm{D}} \left(\int_{B_\tau(x)} |f(x)-f(y)|^\alpha \, dy \right)^{\frac{\beta}{\alpha}} dx \right)^{\frac{1}{\beta}}. \tag{21}$$

4.2. Particular cases.

1. If $\alpha = 1$ and $\beta = 1$,

$$\mathrm{S}_\tau^{(1,1)}(f) = \frac{1}{\operatorname{Vol}_N(\mathrm{D})\,(2\tau)^N} \int_{\mathrm{D}} \int_{B_\tau(x)} |f(x)-f(y)| \, dy\, dx. \tag{22}$$

This is the arithmetic mean of all values $|f(x)-f(y)|$ over $\mathcal{D}_\tau$.

2. If $\alpha = 1$ and $\beta = \infty$,

$$\mathrm{S}_\tau^{(1,\infty)}(f) = (2\tau)^{-N} \sup_{x\in\mathrm{D}} \left(\int_{B_\tau(x)} |f(x)-f(y)| \, dy \right). \tag{23}$$

In this formula, the local arithmetic mean of values $|f(x)-f(y)|$ is evaluated first; then the supremum is taken over x.

3. If $\alpha = \infty$ and $\beta = 1$,

$$\mathrm{S}_\tau^{(\infty,1)}(f) = \frac{1}{\operatorname{Vol}_N(\mathrm{D})} \int_{\mathrm{D}} \left(\sup_{y\in B_\tau(x)} |f(x)-f(y)| \right) dx. \tag{24}$$

For every x, the largest distance between $f(x)$ and $f(y)$ is calculated for y near x. This gives an evaluation of the *oscillation* of f at x. Then the arithmetic mean of oscillations is taken over D. The usual definition for the τ-oscillations of f at x is as follows:

$$\operatorname{osc}_\tau(f;x) = \sup\left\{ f(y) - f(y') \,/\, \|x-y\| \leq \tau,\ \|x-y'\| \leq \tau \right\}.$$

Using the triangle inequality, we get

$$\frac{1}{2}\operatorname{osc}_\tau(f;x) \leq \sup_{y\in B_\tau(x)} |f(x)-f(y)| \leq \operatorname{osc}_\tau(f;x). \tag{25}$$

The integral of τ-oscillations has been studied in [11, 3, 4] and in a number of subsequent publications under the name of the *τ-variation* of f:

$$V_\tau(f) = \int_{\mathrm{D}} \operatorname{osc}_\tau(f;x)\,dx. \tag{26}$$

From (24), (25), and (26), we deduce that $\mathrm{S}_\tau^{(\infty,1)}(f)$ is equivalent as τ tends to 0 to the τ-variation:

$$\frac{1}{2\mathrm{Vol}_N(\mathrm{D})}\,V_\tau(f) \leq \mathrm{S}_\tau^{(\infty,1)}(f) \leq \frac{1}{\mathrm{Vol}_N(\mathrm{D})}\,V_\tau(f). \tag{27}$$

We will see (equation (42)) how the order of growth to 0 of V_τ is directly related to the *fractal dimension* of the graph of f. The same relationship stands for $\mathrm{S}_\tau^{(\infty,1)}$.

4. If $\alpha = \beta = +\infty$,

$$\begin{aligned}\mathrm{S}_\tau^{(\infty,\infty)}(f) &= \sup_{x\in\mathrm{D}}\left(\sup_{y\in B_\tau(x)} |f(x)-f(y)|\right)\\ &= \sup_{(x,y)\in\mathcal{D}_\tau} |f(x)-f(y)|.\end{aligned} \tag{28}$$

This is the maximum τ-oscillation of f over D.

5. If $0 < \alpha = \beta < +\infty$,

$$\mathrm{S}_\tau^{(\alpha,\alpha)}(f) = \left(\frac{1}{\mathrm{Vol}_{2N}(\mathcal{D}_\tau)}\int_{\mathcal{D}_\tau} |f(x)-f(y)|^\alpha\,dy\,dx\right)^{\frac{1}{\alpha}}. \tag{29}$$

This may be written as

$$\mathrm{S}_\tau^{(\alpha,\alpha)}(f) = L^\alpha(F;\mathcal{D}_\tau). \tag{30}$$

Formula (29) is a continuous form of the discrete Kolmogorov means

$$\left[\frac{2}{n(n+1)}\sum_{1\leq j<i}^{n} |x_i - x_j|^\alpha\right]^{\frac{1}{\alpha}},$$

where $x_1,\dots,x_n$ are n real numbers. These mean values lead to the notion of *transfinite diameter* (in the sense of Pólya and Szegö). For a full account of these diameters, see [6].

When $\alpha = 2$, formula (29) is a *standard deviation* between $f(x)$ and $f(y)$ values over the domain $\mathcal{D}_\tau$.

4.3. Properties. Recall that f belongs to $\mathcal{C}(\mathrm{D}(\tau_0))$, $\alpha \in \,]0,+\infty]$, and $\beta \in \,]0,+\infty]$. Properties 1–5 of $\mathrm{S}_\tau^{(\alpha,\beta)}$ come directly from those of $\mathcal{L}^{(\alpha,\beta)}$ (section 3.2). We will give a proof only for property 6. In the particular case of the variation V_τ, properties 1–3 (which characterize a norm) can be found in [12].

THEOREM 4.1. *The operators* $\mathrm{S}_\tau^{(\alpha,\beta)}$ *have the following properties:*

1. $\mathrm{S}_\tau^{(\alpha,\beta)}(f) = 0 \iff f$ *is constant over* D.
2. *For all a,* $\mathrm{S}_\tau^{(\alpha,\beta)}(a\,f) = |a|\,\mathrm{S}_\tau^{(\alpha,\beta)}(f)$.
3. *If* $f_1,\ f_2 \in \mathcal{C}(\mathrm{D}(\tau_0))$, $\alpha \in [1,+\infty]$, *and* $\beta \in [1,+\infty]$, *then*

$$\mathrm{S}_\tau^{(\alpha,\beta)}(f_1+f_2) \leq \mathrm{S}_\tau^{(\alpha,\beta)}(f_1) + \mathrm{S}_\tau^{(\alpha,\beta)}(f_2). \tag{31}$$

4. *Let us fix* $f \in \mathcal{C}(\mathrm{D}(\tau_0))$. $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ *is an increasing and continuous function with respect to the variables* α *and* β *over* $]0,+\infty]$.

5. *Inequalities* (13)–(16) *are still true if we replace* $\mathcal{L}^{(\alpha,\beta)}$ *by* $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ *and* $\|f\|$ *by* $\mathrm{S}_\tau^{(\infty,\infty)}$. *We do not present these new formulas.*

6. *For all* $(\alpha,\beta) \in]0,+\infty] \times]0,+\infty]$,

$$\lim_{\tau\to 0} \mathrm{S}_\tau^{(\alpha,\beta)}(f) = 0. \tag{32}$$

Proof. Since f is uniformly continuous over the compact set $\mathrm{D}(\tau_0)$, the function $\sup_{\mathcal{D}_\tau} |f(x) - f(y)|$ tends to 0 with τ. Hence (32) is true for $\alpha = \beta = +\infty$. When α and β are real numbers, use the inequality $\mathrm{S}_\tau^{(\alpha,\beta)}(f) \leq \mathrm{S}_\tau^{(\infty,\infty)}(f)$. □

4.4. Behavior in the neighborhood of 0. When τ tends to 0, the order of growth of $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ depends on f, α and β. If the domain D is convex, and $\alpha \geq 1$, $\beta \geq 1$, then we may show that this function does not tend to 0 faster than τ.

THEOREM 4.2. *Let* D *be a compact, convex subset of* R^N, $\tau_0 > 0$ *be a real number, and* f *be defined, continuous on* $\mathrm{D}(\tau_0)$, *and nonconstant on* D. *We can find a constant* $c > 0$ (*depending only on* D *and* f) *such that for all* $(\alpha,\beta) \in [1,+\infty] \times [1,+\infty]$,

$$\liminf_{\tau\to 0} \frac{1}{\tau} \mathrm{S}_\tau^{(\alpha,\beta)}(f) \geq c. \tag{33}$$

Since $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ is increasing with respect to α and β (property 4 of Theorem 4.1), it suffices to verify (33) when $\alpha = \beta = 1$.

Since f is uniformly continuous over the compact set $\mathrm{D}(\tau_0)$, for all $\epsilon > 0$, there exists $\tau_1(\epsilon) > 0$, smaller than τ_0, such that

$$\tau \leq \tau_1(\epsilon) \Longrightarrow \forall x \in \mathrm{D},\, \forall y \in B_\tau(x) \,:\, |f(x) - f(y)| \leq \epsilon. \tag{34}$$

To prove the theorem, we will find two real numbers $c_1 > 0$ and c_2 such that

$$\tau \leq \tau_1(\epsilon) \Longrightarrow \int_{\mathcal{D}_\tau} |f(x) - f(y)|\, dx\, dy \geq (c_1 - \epsilon\, c_2)\, \tau^{N+1}. \tag{35}$$

Using (22), the constant c of formula (33) is then equal to

$$c = \frac{c_1}{2^N \operatorname{Vol}_N(\mathrm{D})}.$$

The cases where $N = 1$ and $N \geq 2$ are studied separately.

Proof for $N = 1$. D is a closed interval $[a,b]$ on the line and f is defined and continuous on $[a - \tau_0, b + \tau_0]$. Let

$$c_1 = \sup_{x\in[a,b]} f(x) - \inf_{x\in[a,b]} f(x).$$

This number is strictly positive. There exist two points $y_1 < y_2$ of $[a,b]$ such that

$$|f(y_1) - f(y_2)| = c_1.$$

We can write

$$\int_{\mathcal{D}_\tau} |f(x) - f(y)|\, dx\, dy = \int_{-\tau}^{\tau} \left(\int_a^b |f(x) - f(x+u)|\, dx \right) du,$$

where

$$\int_a^b |f(x)-f(x+u)|\,dx \geq \left|\int_{y_1}^{y_2} (f(x)-f(x+u))\,dx\right|$$
$$= \left|\int_{y_1}^{y_1+u} f(x)\,dx - \int_{y_2}^{y_2+u} f(x)\,dx\right|.$$

Since the inequality $|\int_{y_1}^{y_1+u} f(x)\,dx - u\,f(y_1)| \leq u\,\epsilon$ is also valid near y_2, we get

$$\int_a^b |f(x)-f(x+u)|\,dx \geq u(c_1 - 2\,\epsilon). \tag{36}$$

Finally, by integrating with respect to u on $[-\tau,\tau]$,

$$\int_{\mathcal{D}_\tau} |f(x)-f(y)|\,dx\,dy \geq (c_1 - 2\,\epsilon)\tau^2.$$

This is formula (35), with $N=1$, $c_1 = \sup_{x\in[a,b]} f(x) - \inf_{x\in[a,b]} f(x)$, and $c_2 = 2$. The constant c of Theorem 4.2 is

$$c = \frac{1}{2(b-a)}\left(\sup_{x\in[a,b]} f(x) - \inf_{x\in[a,b]} f(x)\right). \qquad \square$$

Proof for $N \geq 2$. Let us use the Euclidean norm, defined as

$$\|(x_1,\ldots,x_N\|_2 = \sqrt{x_1^2+\cdots+x_N^2}.$$

The Euclidean ball of center x and radius τ is denoted by $B_\tau^{(2)}(x)$. When $x \in \mathrm{D}$, this ball is included in $\mathrm{D}(\tau)$. To obtain (35), it suffices to find c_1 and c_2 such that

$$\tau \leq \tau_1(\epsilon) \Longrightarrow \int_{\mathrm{D}} \int_{B_\tau^{(2)}(0)} |f(x)-f(x+u)|\,du\,dx \geq (c_1 - \epsilon\,c_2)\,\tau^{N+1} \tag{37}$$

with $\tau_1(\epsilon)$ as in (34).

For any u in R^N, let us use the spherical coordinates $u=(\rho,\theta)$, where $\rho = \|u\|_2 \geq 0$ and $\theta = (\theta_1,\ldots,\theta_{N-1})$ belongs to a Cartesian product of intervals denoted by Θ. The Jacobian of the change of coordinates from Cartesian to spherical is $J(\theta) = \rho^{N-1}\,a(\theta)$ with $a(\theta) > 0$ almost everywhere on Θ. The integral $\int_\Theta a(\theta)\,d\theta$ is equal to $2^{N-1}\pi$. It is the $(N-1)$-dimensional volume of the boundary of the unit ball $B_1^{(2)}(0)$. Define

$$I(u) = \int_{\mathrm{D}} |f(x)-f(x+u)|\,dx.$$

The proof is in two parts:

(a) We will first show that there exist two positive functions $c_1(\theta)$ and $c_2(\theta)$ defined on Θ such that

$$\int_\Theta c_1(\theta)\,a(\theta)\,d\theta > 0, \qquad \int_\Theta c_2(\theta)\,a(\theta)\,d\theta < +\infty, \tag{38}$$

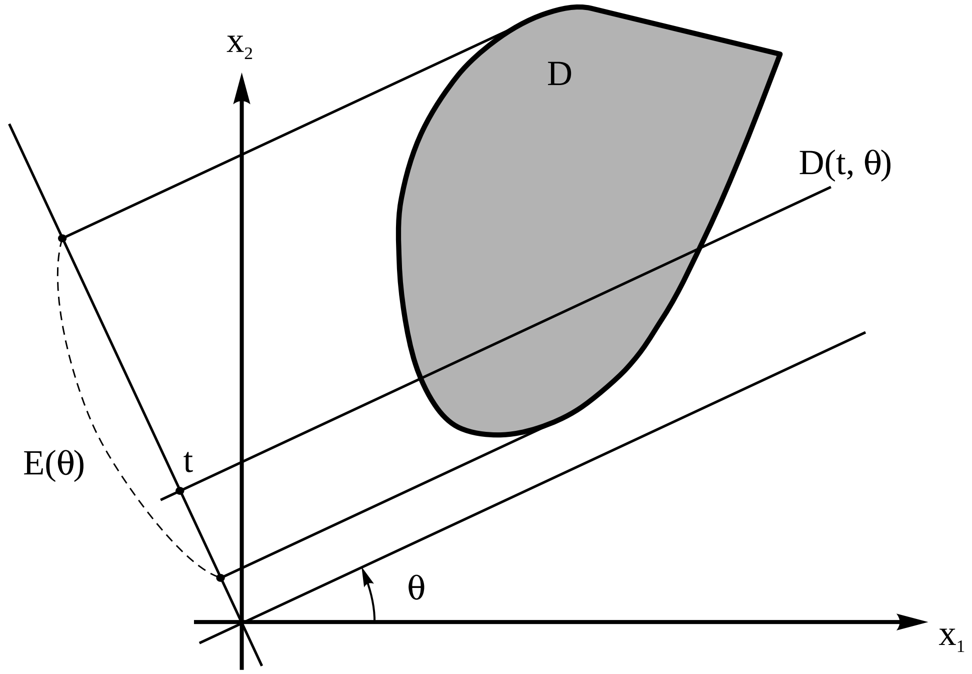


FIG. 1. *The projection $E(\theta)$ and the line $D(t,\theta)$ in the case where $N=2$.*

and for all $u = (\rho, \theta)$,

$$I(u) \geq \rho(c_1(\theta) - \epsilon\, c_2(\theta)). \tag{39}$$

Let us fix u, such that $\rho \neq 0$. Let us denote the following:

$P(\theta)$ is the hyperplane passing through 0 perpendicular to u.

$E(\theta)$ is the orthogonal projection of D over $P(\theta)$.

For all $t \in E(\theta)$, $D(t,\theta)$ is the line passing through t, parallel to u. Since D is convex, the intersection of the line with D is a segment. (See Figure 1.)

$c(t,\theta) = \sup_{x\in \mathrm{D}\cap D(t,\theta)} f(x) - \inf_{x\in \mathrm{D}\cap D(t,\theta)} f(x)$ is the total oscillation of f on $\mathrm{D}\cap D(t,\theta)$. It is a continuous function of (t,θ).

Using a linear, orthogonal change of variables, we can use the case where $N=1$ and (36) to get

$$\int_{D(t,\theta)} |f(x) - f(x+u)|\, dx \geq \rho(c(t,\theta) - 2\epsilon).$$

Integrating with respect to t over $E(\theta)$,

$$I(u) \geq \rho\left(\int_{E(\theta)} c(t,\theta)\, dt - 2\,\epsilon\, \mathrm{Vol}_{N-1}(E(\theta))\right).$$

This proves (39) with

$$c_1(\theta) = \int_{E(\theta)} c(t,\theta)\, dt \tag{40}$$

and

$$c_2(\theta) = 2\,\mathrm{Vol}_{N-1}(E(\theta)).$$

Let us now verify the inequalities in (38). Since f is nonconstant over D, there exists a pair (t_0, θ_0) such that $c(t_0, \theta_0) > 0$. By continuity, $c(t, \theta)$ is strictly positive in a neighborhood of (t_0, θ_0). Also, $a(\theta)$ is positive, continuous, and almost everywhere nonzero. This implies

$$\int_\Theta \int_{E(\theta)} c(t,\theta)\, a(\theta)\, dt\, d\theta > 0.$$

Hence the first inequality stands. The second follows from the boundedness of D.

(b) Now we use (39) to prove (37)

$$\begin{aligned}\int_{\mathrm{D}} \int_{B_\tau^{(2)}(0)} &|f(x) - f(x+u)|\, du\, dx \\ &\geq \int_{B_\tau^{(2)}(0)} \rho(c_1(\theta) - \epsilon\, c_2(\theta))\, J(\theta)\, d\theta\, d\rho \\ &= \int_0^\tau \rho^N\, d\rho \int_\Theta (c_1(\theta) - \epsilon\, c_2(\theta))\, a(\theta)\, d\theta \\ &= \frac{\tau^{N+1}}{N+1} \left(\int_\Theta c_1(\theta)\, a(\theta)\, d\theta - \epsilon \int_\Theta c_2(\theta)\, a(\theta)\, d\theta \right).\end{aligned}$$

To obtain (37), it suffices to choose

$$c_1 = \frac{1}{N+1} \int_\Theta c_1(\theta)\, a(\theta)\, d\theta, \qquad c_2 = \frac{1}{N+1} \int_\Theta c_2(\theta)\, a(\theta)\, d\theta.$$

From (38), we know that $c_1 > 0$ and $c_2 < +\infty$. The constant c of Theorem 4.2 is

$$c = \frac{1}{(N+1)2^N \mathrm{Vol}_N(\mathrm{D})} \int_\Theta \int_{E(\theta)} c(t,\theta)\, a(\theta)\, dt\, d\theta. \qquad \square \tag{41}$$

5. Critical exponents.

5.1. Orders of growth to 0. We keep the notations of section 4. Let us consider a nonconstant function f in $\mathcal{C}(\mathrm{D}(\tau_0))$. Its graph G_f over D is

$$G_f = \{(x, f(x)) \,/\, x \in \mathrm{D}\} \subset R^{N+1}.$$

When the variation $V_\tau(f)$ tends slowly to 0, G_f shows some irregularities at all scales. Such a graph may be called a *fractal* in a weak sense [12]. Let us denote by $\mathrm{Dim}(G_f)$ the fractal dimension. It is a global *irregularity index* which is in direct relationship with the order of growth of $V_\tau(f)$ near 0 [11]. We get

$$\mathrm{Dim}(G_f) = \limsup_{\tau \to 0} \left(N + 1 - \frac{\log V_\tau(f)}{\log \tau} \right). \tag{42}$$

Using (27), this formula is still true after replacing $V_\tau(f)$ by $\mathrm{S}_\tau^{(\infty,1)}(f)$. Analoguously, any of the functions $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ can measure the irregularity of f in a certain sense and gives raise to an irregularity index as follows:

$$\Delta^{(\alpha,\beta)}(f) = \limsup_{\tau \to 0} \left(N + 1 - \frac{\log \mathrm{S}_\tau^{(\alpha,\beta)}(f)}{\log \tau} \right). \tag{43}$$

In this formula, $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$ tends to 0 with τ (see (32)).

5.2. Properties. The properties of operators $\mathrm{S}_\tau^{(\alpha,\beta)}$ have the following consequences:

THEOREM 5.1. *Let* D *be a compact, convex subset of* R^N. *The indices* $\Delta^{(\alpha,\beta)}$ *have the following properties:*

1. *For all* $(\alpha,\beta) \in [1,+\infty] \times [1,+\infty]$,

$$N \le \Delta^{(\alpha,\beta)}(f) \le N+1. \tag{44}$$

2. $\Delta^{(\alpha,\beta)}(f)$ *is increasing with respect to each of the variables* α *and* β.

3. *Here are some relationships between* $\Delta^{(\alpha,\beta)}(f)$ *and* $\Delta^{(\gamma,\delta)}(f)$ *for all strictly positive real numbers* α, β, γ, *and* δ:

If $\gamma \le \alpha$,

$$\Delta^{(\gamma,\beta)}(f) \le \Delta^{(\alpha,\beta)}(f) \le \frac{\gamma}{\alpha}\Delta^{(\gamma,\frac{\beta\gamma}{\alpha})}(f) + \left(1-\frac{\gamma}{\alpha}\right)\Delta^{(\infty,\infty)}(f). \tag{45}$$

If $\delta \le \beta$,

$$\Delta^{(\alpha,\delta)}(f) \le \Delta^{(\alpha,\beta)}(f) \le \frac{\delta}{\beta}\Delta^{(\alpha,\delta)}(f) + \left(1-\frac{\delta}{\beta}\right)\Delta^{(\infty,\infty)}(f). \tag{46}$$

If $\gamma \le \alpha$ *and* $\delta \le \beta$,

$$\Delta^{(\gamma,\delta)}(f) \le \Delta^{(\alpha,\beta)}(f) \le \frac{\gamma\delta}{\alpha\beta}\Delta^{(\gamma,\frac{\gamma\delta}{\alpha})}(f) + \left(1-\frac{\gamma\delta}{\alpha\beta}\right)\Delta^{(\infty,\infty)}(f). \tag{47}$$

Finally, if $\gamma \le \alpha$ *and* $\beta \le \delta$,

$$\begin{aligned}\Delta^{(\gamma,\delta)}(f) &\le \frac{\beta}{\delta}\Delta^{(\alpha,\beta)}(f) + \left(1-\frac{\beta}{\delta}\right)\Delta^{(\infty,\infty)}(f)\\ &\le \frac{\beta\gamma}{\alpha\delta}\Delta^{(\gamma,\frac{\gamma\delta}{\alpha})}(f) + \left(1-\frac{\beta\gamma}{\alpha\delta}\right)\Delta^{(\infty,\infty)}(f).\end{aligned} \tag{48}$$

4. *If* α, β, γ, *and* δ *belong to* $[1,+\infty)$,

$$|\Delta^{(\alpha,\beta)}(f) - \Delta^{(\gamma,\delta)}(f)| \le 1 - \frac{\min\{\alpha,\gamma\}\min\{\beta,\delta\}}{\max\{\alpha,\gamma\}\max\{\beta,\delta\}}. \tag{49}$$

In particular, this shows that $\Delta^{(\alpha,\beta)}(f)$ *is continuous with respect to the two variables* α *and* β *over* $[1,+\infty) \times [1,+\infty)$.

Proof.

(a) Let us verify (44). The right-hand side inequality comes from (32). Formula (33) implies

$$\liminf\left(N+1-\frac{\log \mathrm{S}_\tau^{(\alpha,\beta)}(f)}{\log\tau}\right) \ge \lim\left(N-\frac{\log c}{\log\tau}\right) = N.$$

Therefore, $\Delta^{(\alpha,\beta)}(f) \ge N$.

(b) Since $\mathrm{S}_\tau^{(\alpha,\beta)}$ is increasing, so is $\Delta^{(\alpha,\beta)}$.

(c) Formulas (45)–(48) can be obtained from (13)–(16). Let us assume that $\tau < 1$. Formula (13) gives

$$-\frac{\log \mathrm{S}_\tau^{(\gamma,\beta)}}{\log\tau} \le -\frac{\log \mathrm{S}_\tau^{(\alpha,\beta)}}{\log\tau} \le \frac{\gamma}{\alpha}\left(-\frac{\log \mathrm{S}_\tau^{(\gamma,\frac{\beta\gamma}{\alpha})}}{\log\tau}\right) + \left(1-\frac{\gamma}{\alpha}\right)\left(-\frac{\log \mathrm{S}_\tau^{(\infty,\infty)}}{\log\tau}\right).$$

Now use the inequality

$$\limsup_{\tau\to 0}(g_1(\tau)+g_2(\tau)) \le \limsup_{\tau\to 0} g_1(\tau) + \limsup_{\tau\to 0} g_2(\tau),$$

true for any g_1 and g_2, to get (45).

(d) Formula (49) is implied by the increasing property of $\Delta^{(\alpha,\beta)}$ and by formulas (44), (47), and (48).

If $\gamma \le \alpha$ and $\delta \le \beta$, (47) implies

$$\begin{aligned} 0 \le \Delta^{(\alpha,\beta)} - \Delta^{(\gamma,\delta)} &\le \left(1-\frac{\gamma\delta}{\alpha\beta}\right)(\Delta^{(\infty,\infty)} - \Delta^{(\gamma,\delta)}) \\ &\le 1-\frac{\gamma\delta}{\alpha\beta}. \end{aligned}$$

If $\gamma \le \alpha$ and $\beta \le \delta$, (48) implies

$$\begin{aligned} 0 &\le \Delta^{(\alpha,\beta)} - \Delta^{(\gamma,\delta)} + \left(1-\frac{\beta}{\gamma}\right)(\Delta^{(\infty,\infty)} - \Delta^{(\alpha,\beta)}) \\ &\le \left(1-\frac{\beta\gamma}{\alpha\delta}\right)(\Delta^{(\infty,\infty)} - \Delta^{(\gamma,\delta)}). \end{aligned}$$

Since $\beta\gamma/\alpha\delta \le \beta/\delta$, we get

$$|\Delta^{(\alpha,\beta)} - \Delta^{(\gamma,\delta)}| \le 1-\frac{\beta\gamma}{\alpha\delta}.$$

Similar results are obtained in the other cases. □

6. The study of a nowhere-differentiable function.

6.1. A function defined as a series. Let f be defined on R as follows:

$$f(x) = \sum_{n=0}^{+\infty} 2^{-nH} g(2^n x + \phi_n), \tag{50}$$

where $0 < H < 1$, ϕ_n are real numbers, and g is continuous and periodic with periodicity 1. Then f has also period 1. In this series, amplitudes tend to 0 more slowly than the periods and hence the local oscillations of f. The convergence is absolute so that f is continuous. The phases ϕ_n are here only to help give a more "natural" look to the graph of f for a model of a rough profile, for example [2].

It is well known that such functions are nowhere differentiable and that their graph is a fractal curve. Following the choice of g, we can obtain by this method a Weierstrass function or a Knopp–Takagi function (see [7] for some historical references; see also [9]). When g shows some kind of regularity (when it is continuously differentiable, for example), the fractal dimension of G_f is $2-H$. Finding a general condition on g in order to get this particular result is a difficult task. See, for example, [8]. In this section, we will assume that g verifies one more condition to simplify the calculations of $\mathrm{S}_\tau^{(\alpha,\beta)}(f)$. Using this assumption, we will be able to establish relationships between $\Delta^{(\alpha,\beta)}(f)$ and $\Delta^{(\alpha,\beta)}(g)$, and we will show how to get a fractal dimension different from $2-H$.

6.1.1. Definition domain. Functions g and f are defined on R, but the domain D must be a compact set. We take $\mathrm{D} = [0,1]$ and

$$\mathcal{D}_\tau = \{(x,y) \in R^2 \,/\, x \in [0,1],\ \|x-y\| \leq \tau\}.$$

6.1.2. Notations. Let $f_1(\tau)$ and $f_2(\tau)$ be two positive functions defined in a neighborhood of 0 that tends to 0 with τ. We write

$$\begin{aligned} f_1 \simeq f_2 &\quad \text{if } 0 < \liminf_{\tau\to 0} \frac{f_1(\tau)}{f_2(\tau)} \leq \limsup_{\tau\to 0} \frac{f_1(\tau)}{f_2(\tau)} < +\infty,\\ f_1 \preceq f_2 &\quad \text{if } \limsup_{\tau\to 0} \frac{f_1(\tau)}{f_2(\tau)} < +\infty,\\ f_1 \prec f_2 &\quad \text{if } \lim_{\tau\to 0} \frac{f_1(\tau)}{f_2(\tau)} = 0. \end{aligned}$$

The symbols $\succeq$ and $\succ$ denote the converse relations. We recall that f_1 is *Hölderian with exponent* H at x if $|f_1(x+\tau) - f_1(x)| \preceq \tau^H$. When f_1 is defined on D, it is *uniformly Hölderian with exponent* H on D if there exists a constant C such that for all $x \in \mathrm{D}$ and $y \in \mathrm{D}$, $|f_1(x) - f_1(y)| \leq C|x-y|^H$. An equivalent condition is $\mathrm{S}_\tau^{(\infty,\infty)}(f_1) \preceq \tau^H$.

6.2. Function $g(x)$. We assume that g verifies the following condition:

There exists a constant c such that for all x,

$$g(x) + g\left(x + \frac{1}{2}\right) = c. \tag{51}$$

This constant then is equal to $g(0) + g(1/2) = 2\int_0^1 g(x)\,dx$.

Function g is therefore completely determined on the interval $[0,1/2]$. It may be extended to $]1/2,1]$ using (51) and then to R by periodicity.

6.2.1. Examples.

1. If $g(x) = \cos(2\pi x)$, f is a *Weierstrass function* (see Figure 2). Function g is continuously differentiable, and for all α and β in $[1,+\infty]$, $\mathrm{S}_\tau^{(\alpha,\beta)}(g) \simeq \tau$ in the neighborhhood of 0.

2. If $g(x) = 2x$ on $[0,1/2]$, f is a *Knopp–Takagi function* [7, 5] (see Figure 3). Function g is not everywhere differentiable but rather uniformly Hölderian with exponent 1. For all α and β in $[1,+\infty]$, $\mathrm{S}_\tau^{(\alpha,\beta)}(g) \simeq \tau$.

3. If $g(x) = x^\gamma$ on $[0,1/2]$, $0 < \gamma < 1$ (see Figure 4), the norms $\mathrm{S}_\tau^{(\alpha,\beta)}(g)$ tend to 0 at a rate depending on β. Using the relation

$$(x+\tau)^\gamma - x^\gamma \simeq \tau x^{\gamma-1}$$

when $x \neq 0$, we can show that

$$\mathrm{S}_\tau^{(\alpha,\beta)}(g) \simeq \tau^{\min\{1,\gamma+1/\beta\}}.$$

This result does not depend on α.

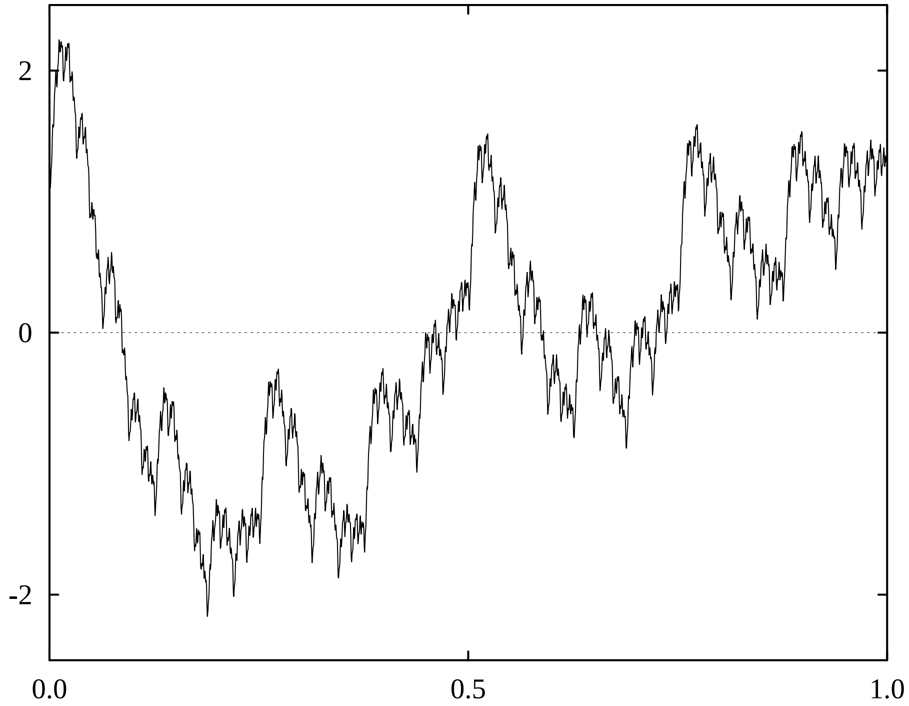


FIG. 2. *A Weierstrass function with* $g(x) = \cos(2\pi x)$*,* $H = 1/2$*, and a sequence* (ϕ_n) *of random numbers.*

6.2.2. Properties. The periodicity of g implies the following.

PROPOSITION 6.1. *For all* $k \geq 1$*,* $\alpha \in [1, +\infty]$*, and* $\beta \in [1, +\infty]$*,*

$$\mathrm{S}_\tau^{(\alpha,\beta)}(g(kx)) = \mathrm{S}_{k\tau}^{(\alpha,\beta)}(g(x)). \tag{52}$$

Proof. By continuity, it suffices to prove (52) for finite α and β. A change of variables gives

$$\mathrm{S}_\tau^{(\alpha,\beta)}(g(kx))^\beta = \frac{1}{(2k\tau)^{\frac{\beta}{\alpha}}} \frac{1}{k} \int_0^k \left(\int_{-k\tau}^{k\tau} |g(y) - g(y+v)|^\alpha \, dv \right)^{\frac{\beta}{\alpha}} dy.$$

By periodicity,

$$\mathrm{S}_\tau^{(\alpha,\beta)}(g(kx))^\beta = \frac{1}{(2k\tau)^{\frac{\beta}{\alpha}}} \int_0^1 \left(\int_{-k\tau}^{k\tau} |g(y) - g(y+v)|^\alpha \, dv \right)^{\frac{\beta}{\alpha}} dy,$$

which gives the result. □

The periodicity of g and assumption (51) imply the following.

PROPOSITION 6.2. *Assume that the constant* c *in* (51) *is* 0*. For any integer* n *and real numbers* ϕ_1 *and* ϕ_2*,*

$$\int_0^1 g(2nx + \phi_1) g(x + \phi_2) \, dx = 0. \tag{53}$$

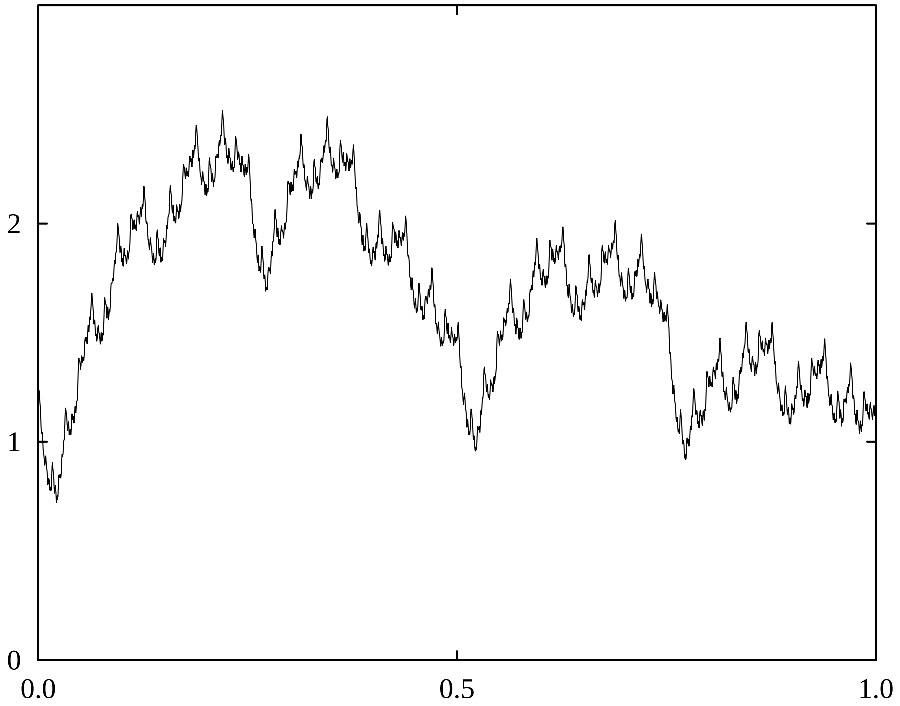


Fig. 3. *A Knopp–Takagi function with $g(x) = 2x$ on $[0, 1/2]$, $H = 1/2$, and a sequence (ϕ_n) of random numbers.*

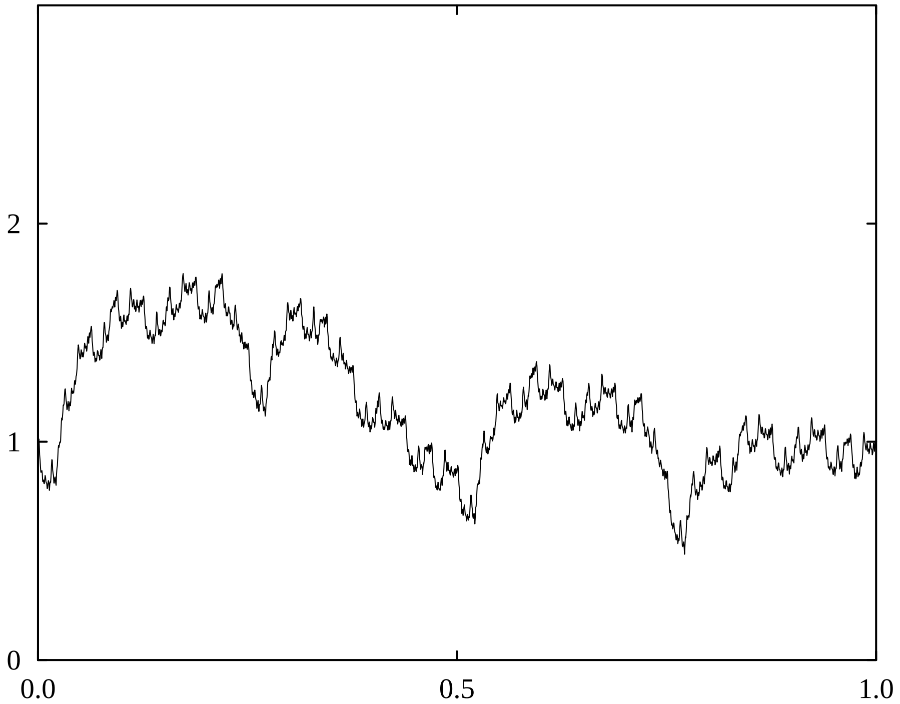


Fig. 4. *Function f obtained with $g(x) = \sqrt{x}$ on $[0, 1/2]$, $H = 1/2$, and a sequence (ϕ_n) of random numbers.*

Proof. The change of variables $x = y + 1/2$ gives

$$\int_{1/2}^{1} g(2nx+\phi_1)g(x+\phi_2)\,dx = \int_0^{1/2} g(2ny+\phi_1+n)g\left(y+\phi_2+\frac{1}{2}\right)\,dy.$$

Since $g(2ny+\phi_1+n) = g(2ny+\phi_1)$ (periodicity) and $g(y+\phi_2+1/2) = -g(y+\phi_2)$ (assumption (51)), we find

$$\int_{1/2}^{1} g(2nx+\phi_1)g(x+\phi_2)\,dx = -\int_0^{1/2} g(2nx+\phi_1)g(x+\phi_2)\,dx.$$

Hence we have the result. □

Corollary 6.3. *Let n and k be two integers, $0 < k < n$, and ϕ_1 and ϕ_2 be two real numbers. Then*

$$\int_{\mathcal{D}_\tau} (g(2^n x+\phi_1) - g(2^n y+\phi_1))(g(2^k x+\phi_2) - g(2^k y+\phi_2))\,dx\,dy = 0. \tag{54}$$

Proof. Without loss of generality, we can asume that the constant c in (51) is 0. Writing the integral as

$$\begin{aligned}\int_{-\tau}^{\tau}\Bigg[\int_0^1 &(g(2^n x+\phi_1)g(2^k x+\phi_2)\\ &- g(2^n x+\phi_1)g(2^k x+\phi_2+2^k u) - g(2^n x+\phi_1+2^n u)g(2^k x+\phi_2)\\ &+ g(2^n x+\phi_1+2^n u)g(2^k x+\phi_2+2^k u))\,dx\Bigg]du,\end{aligned}$$

we see that it suffices to verify the equality

$$\int_0^1 g(2^n x+\phi_1)g(2^k x+\phi_2)\,dx = 0.$$

This last integral can be written as

$$2^{-k}\int_0^{2^k} g(2^{n-k}y+\phi_1)g(y+\phi_2)\,dy$$

or (by periodicity)

$$\int_0^1 g(2^{n-k}y+\phi_1)g(y+\phi_2)\,dy.$$

This is equal to zero (Proposition 6.2). □

6.3. Evaluations of $\boldsymbol{\Delta}^{(\alpha,\beta)}(\boldsymbol{f})$. The previous results will be used for the following.

Theorem 6.4. *For every $\alpha \geq 1$, $\beta \geq 1$, and f as in* (50), *we have*

$$\mathrm{S}_\tau^{(\alpha,\beta)}(f) \succeq \tau^H. \tag{55}$$

Proof. It suffices to prove (55) for $\alpha = \beta = 1$. We take a function $h(x, y)$ defined on $R \times R$ such that $\sup |h| \leq 1$. Let us denote by I the integral $\int_{\mathcal{D}_\tau} |f(x) - f(y)|\, dx\, dy$. Then

$$\begin{aligned} I &\geq \int_{\mathcal{D}_\tau} |h(x, y)(f(x) - f(y))|\, dx\, dy \\ &\geq \left| \int_{\mathcal{D}_\tau} h(x, y)(f(x) - f(y))\, dx\, dy \right| \\ &= \left| \sum_0^{+\infty} 2^{-nH} \int_{\mathcal{D}_\tau} h(x, y)(g(2^n x + \phi_n) - g(2^n y + \phi_n))\, dx\, dy \right|. \end{aligned}$$

If $I_n = \int_{\mathcal{D}_\tau} h(x, y)(g(2^n x + \phi_n) - g(2^n y + \phi_n))\, dx\, dy$,

$$I \geq \sum_0^{+\infty} 2^{-nH} I_n.$$

Let k be the integer such that

$$2^{-k-1} < \tau \leq 2^{-k}$$

and

$$h(x, y) = \frac{1}{2 \sup |g|} (g(2^k x + \phi_k) - g(2^k y + \phi_k)).$$

The condition $\sup |h| \leq 1$ is fulfilled. From Corollary 6.3, $I_n = 0$ for all $n \neq k$. On the other hand,

$$\sqrt{\frac{1}{\tau} I_k} \simeq \mathrm{S}_\tau^{(2,2)}(g(2^k x + \phi_k)).$$

Periodicity implies $\mathrm{S}_\tau^{(2,2)}(g(2^k x + \phi_k)) = \mathrm{S}_\tau^{(2,2)}(g(2^k x))$. From (52), this is the value of $\mathrm{S}_{2^k \tau}^{(2,2)}(g)$. Since $2^k \tau \simeq 1$, $\mathrm{S}_{2^k \tau}^{(2,2)}(g) \simeq \mathrm{S}_1^{(2,2)}(g)$, a constant value. Therefore,

$$\sqrt{\frac{1}{\tau} I_k} \simeq 1.$$

We can deduce that $I_k \simeq \tau$. Since $2^{-kH} \simeq \tau^H$,

$$I \succeq \tau^{1+H}.$$

Finally,

$$\mathrm{S}_\tau^{(1,1)}(f) \succeq \frac{1}{\tau} I \succeq \tau^H. \qquad \square$$

COROLLARY 6.5. *Functions g and f being as before, we have for all $\alpha \geq 1$ and $\beta \geq 1$ that*

$$\Delta^{(\alpha,\beta)}(f) \geq 2 - H. \tag{56}$$

Now let us prove an inequality in the other sense.

Theorem 6.6. *If g is uniformly Hölderian with exponent* 1, *then f is uniformly Holderian with exponent H.*

Proof. Extending the Minkowski inequality to series, we write

$$\mathrm{S}_\tau^{(\infty,\infty)}(f) = \sup_{\mathcal{D}_\tau} |f(x) - f(y)| \leq \sum_0^\infty 2^{-nH}\, \mathrm{S}_\tau^{(\infty,\infty)}(g(2^n x + \phi_n)).$$

Let k be such that $2^{-k-1} < \tau \leq 2^{-k}$. Then

$$\begin{aligned} \mathrm{S}_\tau^{(\infty,\infty)}(g(2^n x + \phi_n)) &= \mathrm{S}_\tau^{(\infty,\infty)}(g(2^n x)) \quad \text{by periodicity} \\ &= \mathrm{S}_{2^n\tau}^{(\infty,\infty)}(g(x)) \quad \text{from (52)} \\ &\preceq \begin{cases} 2^n\tau & \text{if } n \leq k \text{ since } g \text{ is Hölderian,} \\ 1 & \text{if } n > k \text{ since } g \text{ is periodic.} \end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{S}_\tau^{(\infty,\infty)}(f) &\preceq \sum_0^k 2^{-nH} 2^n \tau + \sum_{k+1}^\infty 2^{-nH} \\ &\simeq \tau 2^{k(1-H)} + 2^{-kH} \\ &\simeq \tau^H. \qquad \square \end{aligned}$$

Corollary 6.7. *If g is uniformly Hölderian with exponent* 1, *then for all $\alpha \geq 1$ and $\beta \geq 1$,*

$$\Delta^{(\alpha,\beta)}(f) \leq 2 - H. \tag{57}$$

Notice that we have used the relation $\mathrm{S}_\tau^{(\infty,\infty)}(g) \preceq \tau$ in the proof of Theorem 6.6. Since $H < 1$, this theorem is a particular case of the following.

Theorem 6.8. *Let $\alpha \in [1,\infty]$, $\beta \in [1,\infty]$, and $\omega > 0$ be such that $\mathrm{S}_\tau^{(\alpha,\beta)}(g) \preceq \tau^\omega$. Then*

(i) *If $\omega \neq H$, $\mathrm{S}_\tau^{(\alpha,\beta)}(f) \preceq \tau^{\min\{H,\omega\}}$.*

(ii) *If $\omega = H$, $\mathrm{S}_\tau^{(\alpha,\beta)}(f) \preceq \tau^H |\log \tau|$.*

Proof. Use the inequality

$$\mathrm{S}_\tau^{(\alpha,\beta)}(f) \leq \sum_0^\infty 2^{-nH} \mathrm{S}_\tau^{(\alpha,\beta)}(g(2^n x + \phi_n)),$$

where $\mathrm{S}_\tau^{(\alpha,\beta)}(g(2^n x+\phi_n)) = \mathrm{S}_{2^n\tau}^{(\alpha,\beta)}(g) \preceq (2^n\tau)^\omega$. Let k be such that $2^{-k-1} < \tau \leq 2^{-k}$. We get

$$\sum_0^k 2^{-nH} \mathrm{S}_\tau^{(\alpha,\beta)}(g(2^n x + \phi_n)) \preceq \tau^\omega \sum_0^k 2^{n(\omega - H)}.$$

The right-hand side member has the same order of growth as

$$\begin{array}{ll} \tau^\omega 2^{k(\omega-H)} \simeq \tau^H & \text{if } \omega > H, \\ k\,\tau^\omega \simeq \tau^H |\log \tau| & \text{if } \omega = H, \\ \tau^\omega & \text{if } \omega < H. \end{array}$$

On the other hand, $\mathrm{S}^{(\alpha,\beta)}_{2^n\tau}(g) \preceq 1$ by periodicity, and

$$\sum_{k+1}^{\infty} 2^{-nH}\mathrm{S}^{(\alpha,\beta)}_{\tau}(g(2^n x+\phi_n)) \preceq \sum_{k+1}^{\infty} 2^{-nH} \simeq \tau^H.$$

Gathering these results, the proof is completed. □

Corollary 6.9. *For all $\alpha \geq 1$ and $\beta \geq 1$,*

$$\Delta^{(\alpha,\beta)}(f) \leq \max\{2-H, \Delta^{(\alpha,\beta)}(g)\}. \tag{58}$$

Proof. Let ω be such that $\Delta^{(\alpha,\beta)}(g) \leq 2-\omega$ and $\omega \neq H$. There exists $\tau_1 < 1$ such that

$$\tau \leq \tau_1 \Longrightarrow 2 - \frac{\log \mathrm{S}^{(\alpha,\beta)}_{\tau}(g)}{\log \tau} \leq 2-\omega,$$

i.e., $\mathrm{S}^{(\alpha,\beta)}_{\tau}(g) \leq \tau^{\omega}$. Theorem 6.8 implies $\mathrm{S}^{(\alpha,\beta)}_{\tau}(f) \succeq \tau^{\min\{H,\omega\}}$. Hence

$$\Delta^{(\alpha,\beta)}(f) \leq 2-\min\{H,\omega\}.$$

With ω tending to $2-\Delta^{(\alpha,\beta)}(g)$, we get (58). □

Conjecture. *Let us denote by $\delta^{(\alpha,\beta)}(f)$ the following index:*

$$\delta^{(\alpha,\beta)}(f) = \liminf_{\tau\to 0}\left(2 - \frac{\log \mathrm{S}^{(\alpha,\beta)}_{\tau}}{\log \tau}\right).$$

Then for all $\alpha \geq 1$ and $\beta \geq 1$,

$$\delta^{(\alpha,\beta)}(f) \geq \max\{2-H, \delta^{(\alpha,\beta)}(g)\}. \tag{59}$$

This formula would be the symmetric to (58).

A particular case. This conjecture is true if $\alpha=\beta=2$. To show this, we write

$$\begin{aligned}\mathrm{S}^{(2,2)}_{\tau}(f)^2 &= \sum_{0}^{\infty} 2^{-2nH}\mathrm{S}^{(2,2)}_{\tau}(g(2^n x+\phi_n))^2\\ &= \sum_{0}^{\infty} 2^{-2nH}\mathrm{S}^{(2,2)}_{2^n\tau}(g)^2.\end{aligned}$$

Let ω be such that $\delta^{(2,2)}(g) \geq 2-\omega$. Then $\mathrm{S}^{(2,2)}_{\tau}(g) \succeq \tau^{\omega}$. Let k be the integer such that $2^{-k-1} < \tau \leq 2^{-k}$. Then

$$\mathrm{S}^{(2,2)}_{\tau}(f)^2 \succeq \sum_{0}^{k} 2^{-2nH}(2^n\tau)^{2\omega}.$$

The second member of this relation has the same order of growth as

$$\begin{array}{ll}\tau^{2\omega}2^{2k(\omega-H)} \simeq \tau^{2H} & \text{if } \omega > H,\\ k\,\tau^{2\omega} \simeq \tau^{2H}|\log\tau| & \text{if } \omega = H,\\ \tau^{2\omega} & \text{if } \omega < H.\end{array}$$

Therefore,

$$\mathrm{S}_\tau^{(2,2)}(f) \succeq \begin{cases} \tau^{\min(H,\omega)} & \text{if } \omega \neq H, \\ \tau^H \sqrt{|\log \tau|} & \text{if } \omega = H. \end{cases}$$

This proves that $\delta^{(2,2)}(f) \geq 2 - \min\{H, \omega\}$. Letting ω tend to $2 - \delta^{(2,2)}(g)$, we get (59).

Remarks.

1. Corollary 6.5 confirms the conjecture.
2. Corollaries 6.5 and 6.7 prove that $\Delta^{(\alpha,\beta)}(f) = 2 - H$ for all $\alpha \geq 1$ and $\beta \geq 1$ when g is uniformly Hölderian with exponent 1.

The irregularity indices may take values different from $2 - H$. First, let us prove the following.

Proposition 6.10. *We have*

$$\Delta^{(1,1)}(f) \geq 2\Delta^{(2,2)}(g) - \Delta^{(\infty,\infty)}(g). \tag{60}$$

Proof. Using the same notations as in the proof of Theorem 6.4, recall that

$$I \geq \sum_0^{+\infty} 2^{-nH} I_n.$$

Let us choose

$$h(x, y) = \frac{g(x + \phi_0) - g(y + \phi_0)}{\mathrm{S}_\tau^{(\infty,\infty)}(g)}.$$

We get $I_n = 0$ for all $n \geq 1$, and

$$I_0 \simeq \tau \frac{\mathrm{S}_\tau^{(2,2)}(g)^2}{\mathrm{S}_\tau^{(\infty,\infty)}(g)}.$$

There exists a constant $c_1 > 0$ such that

$$\mathrm{S}_\tau^{(1,1)}(f) \geq c_1 \frac{\mathrm{S}_\tau^{(2,2)}(g)^2}{\mathrm{S}_\tau^{(\infty,\infty)}(g)},$$

which can be written as

$$2 - \frac{\log \mathrm{S}_\tau^{(1,1)}(f)}{\log \tau} \geq 2\left(2 - \frac{\log \mathrm{S}_\tau^{(2,2)}(g)}{\log \tau}\right) - \left(2 - \frac{\mathrm{S}_\tau^{(\infty,\infty)}(g)}{\log \tau}\right) - \frac{\log c_1}{\log \tau}.$$

Formula (60) is obtained by taking the lim sup on both sides of this inequality. □

Corollary 6.11. *If* $\Delta^{(2,2)}(g) = \Delta^{(\infty,\infty)}(g)$*, then for all* $\alpha \geq 1$ *and* $\beta \geq 1$*,*

$$\Delta^{(\alpha,\beta)}(f) = \max\{2 - H, \Delta^{(\infty,\infty)}(g)\}. \tag{61}$$

Proof. With this assumption, the right-hand side member of (60) is $\Delta^{(\infty,\infty)}(g)$. Then $\Delta^{(\alpha,\beta)}(f) \geq \Delta^{(1,1)}(f) \geq \Delta^{(\infty,\infty)}(g)$. The proof is completed with the help of (56) and (58). □

6.4. Some applications.

1. If $g(x) = \cos 2\pi x$, $f(x)$ is a Weierstrass function. Since $g(x)$ is differentiable, Corollaries 6.5 and 6.7 can be used, and

$$\Delta^{(\alpha,\beta)}(f) = 2 - H$$

for all $\alpha \geq 1$ and $\beta \geq 1$. The same result holds for the Knopp–Takagi function, where $g(x) = 2x$ on $[0, 1/2]$. This shows in particular that the fractal dimension of the graph of the Takagi function is $2 - H$, a result that is very easy to prove when the phases ϕ_n are all zero but more difficult otherwise. Another proof in a different context can be found in [1].

2. Let g be the Weierstrass function

$$g(x) = \sum_{0}^{\infty} 2^{-n\omega} \cos(2\pi 2^n x + \phi_n)$$

on $[0, 1/2]$. The parameter ω is in $]0, 1[$. Then $\mathrm{S}_\tau^{(\alpha,\beta)}(g) \simeq \tau^\omega$. Corollary 6.11 lets us conclude that

$$\Delta^{(\alpha,\beta)}(f) = 2 - \min(H, \omega)$$

for any $\alpha \geq 1$ and $\beta \geq 1$.

3. Let $g(x) = \sqrt{x}$ on $[0, 1/2]$. Then

$$\mathrm{S}_\tau^{(\alpha,\beta)}(g) \simeq \tau^{\min(1, \frac{1}{2} + \frac{1}{\beta})}$$

(section 6.2.1). It follows that

$$\Delta^{(\alpha,\beta)}(g) = \max\left\{1, \frac{3}{2} - \frac{1}{\beta}\right\}.$$

(i) If $\beta \leq 2$, then $\Delta^{(\alpha,\beta)}(g) = 1$. From Corollaries 6.5 and 6.7, we deduce that

$$\Delta^{(\alpha,\beta)}(f) = 2 - H$$

for all $\alpha \geq 1$. This is the value of the fractal dimension of G_f in particular.

(ii) If $\beta \geq 2$, $\Delta^{(\alpha,\beta)}(g) = 3/2 - 1/\beta$. If, for example, $H = 3/4$, then Corollary 6.9 gives

$$\Delta^{(\alpha,\beta)}(f) \leq \begin{cases} \frac{5}{4} & \text{if } 2 \leq \beta \leq 4, \\ \frac{3}{2} - \frac{1}{\beta} & \text{if } \beta \geq 4. \end{cases}$$

The equality is conjectural.

REFERENCES

[1] A. Baouche and S. Dubuc, *A unified approach for non differentiable functions*, J. Math. Anal. Appl., 182 (1994), pp. 134–142.

[2] M. V. Berry and Z. V. Lewis, *On the Weierstrass-Mandelbrot fractal function*, Proc. Roy. Soc. London Ser. A, 370 (1980), pp. 459–484.

[3] B. Dubuc and C. Tricot,*Variation d'une fonction et dimension de son graphe*, C. R. Acad. Sci. Paris Sér. I Math., 306 (1988), pp. 531–533.

[4] B. Dubuc, C. Tricot, S. W. Zucker, and C. Roques-Carmes, *Evaluating the fractal dimension of surfaces*, Proc. Roy. Soc. London Ser. A, 425 (1989), pp. 113–127.

[5] M. Hata and M. Yamaguti, *The Takagi function and its generalization*, Japan. J. Appl. Math., 1 (1984), pp. 183–199.

[6] E. Hille, *Methods in Classical and Functional Analysis*, Addison–Wesley, Reading, MA, 1972.

[7] E. W. Hobson, *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, Vol. II, Dover, New York, 1957.

[8] J. L. Kaplan, J. Mallet-Paret, and J. A. Yorke, *The Lyapounov dimension of a nowhere differentiable attracting torus*, Ergodic Theory Dynamical Systems, 4 (1984), pp. 261–281.

[9] B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, San Francisco, 1982.

[10] P. Maragos and F. K. Sun, *Measuring the fractal dimension of signals*, IEEE Trans. Signal Process., 41 (1993), pp. 108–121.

[11] C. Tricot, C. Roques-Carmes, J. F. Quiniou, D. Wehbi, and B. Dubuc, *Evaluation de la dimension fractale d'un graphe*, Rev. Phys. Appl., 23 (1988), pp. 111–124.

[12] C. Tricot, *Courbes et Dimension Fractale*, Springer-Verlag, Berlin, 1993 (in French); *Curves and Fractal Dimension*, Springer-Verlag, Berlin, 1994 (in English).

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 213–232, January 1997

011

INEQUALITIES ON MATRIX-DILATED LITTLEWOOD–PALEY ENERGY FUNCTIONS AND OVERSAMPLED AFFINE OPERATORS*

CHARLES K. CHUI[†] AND XIANLIANG SHI[‡]

Abstract. Affine operators and Littlewood–Paley energy functions with matrix dilations are considered in this paper. Estimates and comparisons of the infimum and supremum measurements of these two operations are derived. These results are applied to the study of affine frames and wavelets. In particular, multivariate matrix-dilated wavelet families are characterized and a matrix-dilation oversampling theorem on preservation of frame-bound ratios is established.

Key words. matrix dilation, Littlewood–Paley, inequalities, oversampling

AMS subject classifications. 42C15, 42C99

PII. S0036141094272009

1. Introduction and results.

The theory of s-dimensional discrete wavelet transform is the mathematical analysis of a family

$$\{\psi_{\ell,b;j,k}\colon\ 1 \le \ell \le L, \quad j \in \mathbb{Z}, \quad k \in \mathbb{Z}^s\} \tag{1.1}$$

of functions generated by some functions $\psi_1, \dots, \psi_L$ in $L^2 := L^2(\mathbb{R}^s)$, in the sense that

$$\psi_{\ell,b;j,k}(x) := b^{s/2} |\det M|^{j/2} \psi_\ell(M^j x - kb), \tag{1.2}$$

where $b > 0$ is the (discretization) sample period and M is a nonsingular $s \times s$ matrix whose eigenvalues $\lambda_1, \dots, \lambda_s$ satisfy

$$|\lambda_\alpha| > 1, \quad \alpha = 1, \dots, s. \tag{1.3}$$

While the special case where $M = 2I_s$, where I_s denotes the s-dimensional identity matrix, is a somewhat straightforward generalization of the univariate theory, this paper is concerned with more general dilation matrices M. The objective of this paper is to establish some inequalities on the Littlewood–Paley energy functions and affine operators and to apply these results to the study of affine frames and wavelets and the boundedness of the Littlewood–Paley g-function operators.

Let us first introduce some notations. For a given family

$$\Psi := \{\psi_1, \dots, \psi_L\} \tag{1.4}$$

*Received by the editors July 29, 1994; accepted for publication September 14, 1995. This research was supported by NSF grant DMS-92-06928, ARO contract DAAH-04-93-G-0047, and Texas Higher Education Coordinating Board grant 999903-054.

http://www.siam.org/journals/sima/28-1/27200.html.

[†] Center for Approximation Theory, Texas A&M University, College Station, TX 77843 (cchui@tamu.edu).

[‡] Department of Mathematics, Texas A&M University, College Station, TX 77843 (xshi@math.tamu.edu).

of functions in L^2, the operator T_b, defined by

$$(T_b f)(\mathbf{x}) := \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z},\mathbf{k}\in\mathbb{Z}^s} \langle f, \psi_{\ell,b;j,\mathbf{k}}\rangle \psi_{\ell,b;j,\mathbf{k}}(\mathbf{x}), \quad f \in L^2, \tag{1.5}$$

will be called an *affine operator*, and the infinite series

$$L_\Psi(\mathbf{x}) := \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} |\widehat{\psi}_\ell((M^T)^j \mathbf{x})|^2 \tag{1.6}$$

will be called the corresponding *Littlewood–Paley energy function* associated with the family Ψ. Here M^T denotes the transpose of M and the Fourier transform $\widehat{\psi}$ of ψ is defined, as usual, by

$$\widehat{\psi}(x) = \int_{\mathbb{R}^s} \psi(t) e^{-ixt} dt.$$

We will also use the standard operator norm notation $\|T\|$ for bounded linear operators T on L^2. It is clear that

$$\|T\| = \sup_{\substack{f,g\in\mathcal{D} \\ \|f\|\le 1, \|g\|\le 1}} \langle Tf, g\rangle, \tag{1.7}$$

where $\|f\|$ denotes the L^2 norm of f and $\mathcal{D}$ denotes the set of all infinitely differentiable and compactly supported functions. Let us also consider the measurement

$$\|T\|^* := \sup_{\|f\|=1} |\langle Tf, f\rangle| \tag{1.8}$$

and observe that $\|T\|^* \le \|T\|$ always holds for any bounded operator on L^2, but the two quantities are different in general. For certain operators, however, such as the affine operators in (1.5) and the Fourier multipliers defined by $\widehat{Tf} = m\widehat{f}$, $m \in L^\infty$, we do have equality.

The main contents of this paper are estimation and comparison of the quantities

$$\begin{cases} \|L_\Psi\|_* := \operatorname{ess\,inf}_{\mathbf{x}} |L_\Psi(\mathbf{x})|, \\ \|L_\Psi\|^* := \operatorname{ess\,sup}_{\mathbf{x}} |L_\Psi(\mathbf{x})|, \\ \|T_b\|_* := \inf_{\|f\|=1} |\langle T_b f, f\rangle|, \quad \text{and} \\ \|T_b\|^* \quad \text{as defined in (1.8)} \end{cases}$$

and applications of these results to the study of affine frames and wavelets.

In the one-dimensional setting with $M = (2)$, $b = 1$, and $L = 1$, we recall from Mallat and Zhong [5] that ψ is called a *dyadic wavelet* if the sequence

$$(W_j f)(x) := f * (2^j \psi(2^j \cdot))(x), \quad j \in \mathbb{Z},$$

satisfies the so-called stability condition

$$A\|f\|^2 \le \sum_{j\in\mathbb{Z}} \|W_j f\|^2 \le B\|f\|^2, \quad f \in L^2,$$

where $0 < A \le B < \infty$ are constants independent of f. This concept is instrumental to the construction of dyadic duals for the recovery of $f \in L^2$ from the sequence $W_j f$, $j \in \mathbb{Z}$. It is clear that ψ is a dyadic wavelet if and only if the corresponding Littlewood–Paley energy function L_ψ defined in (1.6) is bounded both from above and away from zero. In our earlier work [2], we also established the critera for L_ψ to have these boundedness properties, namely, $\psi \in L^2$ with $\int \psi = 0$,

$$\text{ess inf}\left\{|\widehat{\psi}(\omega)|\colon\ \frac{1}{2}a \le |\omega| \le a\right\} > 0$$

for some $a > 0$, and

$$|\psi(x)| \le \Phi(|x|), \quad x \in \mathbb{R},$$

for some nonnegative and nonincreasing function Φ on $(0, \infty)$ that satisfies

$$\Phi(0) + \int_1^\infty \Phi(x)(\ln x)^{1/2} dx < \infty.$$

However, the method used in [2] for establishing this one-variable result does not generalize to the study of the multidimensional setting with a more general dilation matrix M. In sections 2 and 3, we will establish certain upper-bound and lower-bound results in the multidimensional setting with matrix dilations. Assuming that Φ is a nonincreasing continuous function defined on $[0, \infty)$ and satisfies

$$K_\Phi := \int_{\mathbb{R}^s} \Phi(|\mathbf{x}|)\left(1 + \sqrt{\ln^+ |\mathbf{x}|}\right) d\mathbf{x} < \infty, \tag{1.9}$$

we will establish the following in section 2.

Theorem 1. *Let Φ be a nonincreasing function on $[0, \infty)$ that satisfies (1.9) and $\Psi = \{\psi_1, \dots, \psi_L\}$ be a collection of functions in L^2.*

(i) *If $\|T_b\| < \infty$, then*

$$\|L_\Psi\|^* \le \|T_b\|^* = \|T_b\|. \tag{1.10}$$

(ii) *Let Ψ satisfy*

$$|\psi_\ell(\mathbf{x})| \le \Phi(|\mathbf{x}|), \quad \ell = 1, \dots, L. \tag{1.11}$$

Then there exists some positive constant C, independent of Φ, such that

$$\|L_\Psi\|^* \le C K_\Phi^2 \tag{1.12}$$

if and only if

$$\int \psi_\ell(\mathbf{x}) d\mathbf{x} = 0, \quad \ell = 1, \dots, L. \tag{1.13}$$

Furthermore, (1.11) *cannot be replaced by the simple assumption $\psi_\ell \in L^1 \cap L^2$, $\ell = 1, \dots, L$, to ensure $\|L_\Psi\|^* < \infty$.*

For certain more specific dilation matrices M, we can even compare the quantities $\|L_\Psi\|_*$ and $\|T_b\|_*$. More precisely, the following result will be established in section 3.

Theorem 2. *Let $M = \lambda U$, where $\lambda > 1$ and U is a unitary matrix, and assume that $\|T_b\| < \infty$. Then*

$$\|L_\Psi\|_* \geq \|T_b\|_*. \tag{1.14}$$

In section 4, we will compare the values of $\|T_b\|^*$ and $\|T_b\|_*$ with those of their corresponding oversampled operators. In particular, the following result is established.

Theorem 3. *Suppose that M is any nonsingular square matrix with integer entries and with all its eigenvalues satisfying* (1.3). *Let n be any natural number that satisfies*

$$(n, |\det M|) = 1 \tag{1.15}$$

and Ψ be a finite family with corresponding affine operator T_b that satisfies $\|T_b\|^* < \infty$. Then

$$\|T_b\|_* \leq \|T_{b/n}\|_* \leq \|T_{b/n}\|^* \leq \|T_b\|^*. \tag{1.16}$$

Furthermore, (1.16) *does not hold in general without the assumption* (1.15).

Finally, we will discuss, in section 5, an application of these results to the theory of affine frames and wavelets and a study of the boundedness of Littlewood–Paley g-function operators.

2. Proof of Theorem 1.

This section is devoted to the proof of Theorem 1.

2.1. Preliminary results.

To facilitate the proof of the theorem, we need the following three lemmas. First, observe that if M satisfies (1.3), then we have for some positive $J \in \mathbb{Z}$,

$$\theta := \min_{\mathbf{x} \in \sigma} |(M^T)^J \mathbf{x}| > 1, \tag{2.1}$$

where σ denotes the unit sphere

$$\sigma := \{\mathbf{x} \in \mathbb{R}^s \colon\ |\mathbf{x}| = 1\}.$$

By (2.1), we see that σ lies in the bounded component of the complement of its image

$$M^J(\sigma) := \{\mathbf{y} = (M^T)^J \mathbf{x} \colon\ \mathbf{x} \in \sigma\}$$

under $(M^T)^J$.

The following lemma is evident.

Lemma 1. *If M satisfies* (1.3), *then for any $\mathbf{x} \in \mathbb{R}^s \backslash \{\mathbf{0}\}$, there exists some $j_0 \in \mathbb{Z}$ and $\mathbf{x}_0 \in \Omega_J(M)$ such that $M^{j_0 J}\mathbf{x}_0 = \mathbf{x}$, where*

$$\Omega_J(M) := \Omega_J(M)^0 \cup \sigma \tag{2.2}$$

and $\Omega_J(M)^0$ denotes the open region with boundary $\sigma \cup M^J(\sigma)$. Furthermore, the pair $(j_0, \mathbf{x}_0)$ is unique.

In the following, $L^2(Q)$ will denote the space of all square-integrable functions f on a measurable set Q with norm

$$\|f\|_{L^2(Q)} := \left(\int_Q |f(\mathbf{x})|^2 d\mathbf{x} \right)^{1/2}.$$

LEMMA 2. *Let Q be any measurable set in $\mathbb{R}^s$ and $\{f_j\}_{j\in\mathbb{Z}}$ be a sequence in $L^2(Q)$. Then the following two statements are equivalent:*

(i) $\|\sum_j c_j f_j\|^2_{L^2(Q)} \le B\sum_j |c_j|^2,\ \{c_j\}\in\ell^2.$

(ii) $\sum_j |\int_Q f\bar{f}_j|^2 \le B\|f\|^2_{L^2(Q)},\ f\in L^2(Q).$

The following lemma can be found in [7, p. 43].

LEMMA 3. *For any real number $\delta\notin\mathbb{Z}\cup\big(\mathbb{Z}+1/2\big)$, the identity*

$$1-e^{i\delta t}=F_1(\delta)+F_2(\delta,t)+F_3(\delta,t),\quad t\in[-\pi,\pi], \tag{2.3}$$

holds, where

$$F_1(\delta):=\left(1-\frac{\sin\pi\delta}{\pi\delta}\right),$$

$$F_2(\delta,t):=\sum_{k=1}^{\infty}\frac{(-1)^k 2\delta\sin\pi\delta}{\pi(k^2-\delta^2)}\cos kt,$$

and

$$F_3(\delta,t):=i\sum_{k=1}^{\infty}\frac{(-1)^k 2\delta\cos\pi\delta}{\pi\big(\big(k-\frac{1}{2}\big)^2-\delta^2\big)}\sin\left(k-\frac{1}{2}\right)t.$$

2.2. Proof of Theorem 1(i). Let $\|T_b\|<\infty$. Then for any f with $\|f\|=1$, we have

$$\begin{aligned}\langle T_b f,f\rangle &= \frac{b^s}{(2\pi)^{2s}}\sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z},\mathbf{k}\in\mathbf{Z}^s}|\det M|^j\left|\int_{\mathbb{R}^s}\widehat{f}((M^T)^j\mathbf{x})\overline{\widehat{\psi}_\ell(\mathbf{x})}e^{ib\mathbf{kx}}d\mathbf{x}\right|^2\\ &=\sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\frac{|\det M|^j}{b^s}\sum_{\mathbf{k}\in\mathbf{Z}^s}\left|\frac{1}{(2\pi b^{-1})^s}\int_{[0,2\pi b^{-1}]^s}\right.\\ &\qquad\left.\left[\sum_{\mathbf{m}\in\mathbf{Z}^s}\widehat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi}_\ell(x+2\pi b^{-1}\mathbf{m})}e^{ib\mathbf{kx}}\right]^2 d\mathbf{x}\right|.\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\frac{|\det M|^j}{(2\pi)^s}\int_{[0,2\pi b^{-1}]^s}\left|\sum_{\mathbf{m}\in\mathbf{Z}^s}\widehat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi}_\ell(\mathbf{x}+2\pi b^{-1}\mathbf{m})}\right|^2 d\mathbf{x}\\ &\quad\le\|T_b\|^*.\end{aligned}$$

On the other hand, by using the notation $\mathbf{e}:=(1,\dots,1)$ and considering any $\mathbf{x}_0\in\mathbb{R}^s\backslash\{\mathbf{0}\}$, we have for any positive integer J' that

$$\begin{aligned}&\sum_{\ell=1}^{L}\sum_{j=-J'}^{J'}\frac{|\det M|^j}{(2\pi)^s}\int_{[(M^T)^{-j}\mathbf{x}_0-\pi_b^{-1}\mathbf{e},(M^T)^{-j}\mathbf{x}_0+\pi b^{-1}\mathbf{e}]}\\ &\left|\sum_{m\in\mathbf{Z}^s}\widehat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi}_\ell(\mathbf{x}+2\pi b^{-1}\mathbf{m})}\right|^2 d\mathbf{x}\le\|T_b\|^*.\end{aligned} \tag{2.4}$$

Now if we set $\widehat{f}(\mathbf{x}) = (\pi\varepsilon^{-1})^{s/2}\chi_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}(\mathbf{x})$ in (2.4) and take the limit as $\varepsilon \to 0$ and $J' \to \infty$ consecutively, we arrive at (1.10). This completes the proof of Theorem 1(i).

2.3. Proof of Theorem 1(ii). Assume that (1.11) holds. We will first show that

$$\sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}|\widehat{\psi}_\ell((M^T)^{Jj}\mathbf{x})|^2 \le C\left\{\int_{\mathbb{R}^s}\Phi(|\mathbf{t}|)\big(1+\sqrt{\ln^+|\mathbf{t}|}\big)d\mathbf{t}\right\}^2, \quad \mathbf{x}\neq 0. \tag{2.5}$$

Observe that since

$$\sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}|\widehat{\psi}_\ell((M^T)^{Jj}((M^T)^{J}\mathbf{x}))|^2 = \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}|\widehat{\psi}_\ell((M^T)^{Jj}\mathbf{x})|^2,$$

an application of Lemma 1 shows that (2.5) is valid provided that it holds for any $\mathbf{x}\in\Omega_J(M)$. For this reason, we call $\Omega_J(M)$ a "fundamental region."

By (1.11) and (1.13), we see that

$$\begin{aligned}
|\widehat{\psi}_\ell(M^T)^{Jj}\mathbf{x})| &= \left|\int_{\mathbb{R}^s}\psi_\ell(\mathbf{t})(e^{-i\mathbf{t}(M^T)^{Jj}\mathbf{x}}-1)d\mathbf{t}\right| \\
&\le |(M^T)^{Jj}\mathbf{x}|\int_{|\mathbf{t}|\le|(M^T)^{Jj}\mathbf{x}|^{-1}}\Phi(|\mathbf{t}|)|\mathbf{t}|d\mathbf{t} + 2\int_{|\mathbf{t}|\ge|(M^T)^{Jj}\mathbf{x}|^{-1}}\Phi(|\mathbf{t}|)d\mathbf{t} \\
&=: K_{1,j}+K_{2,j},
\end{aligned}$$

and this gives

$$\begin{aligned}
P_1(\mathbf{x}) &:= \sum_{\ell=1}^{L}\sum_{j=-\infty}^{-1}|\widehat{\psi}_\ell((M^T)^{Jj}\mathbf{x})|^2 \\
&\le 2L\sum_{j=1}^{\infty}|K_{1,-j}|^2 + 2L\sum_{j=1}^{\infty}|K_{2,-j}|^2 \\
&=: 2P_{1,1}(\mathbf{x}) + 2P_{1,2}(\mathbf{x}).
\end{aligned} \tag{2.6}$$

Next, consider

$$\mu_0 := \int_{|\mathbf{t}|\le|(M^T)^{-J}\mathbf{x}|^{-1}}\Phi(|\mathbf{t}|)|\mathbf{t}|d\mathbf{t}$$

and

$$\mu_k := \int_{|(M^T)^{-Jk}\mathbf{x}|^{-1}\le|\mathbf{t}|\le|(M^T)^{-J(k+1)}\mathbf{x}|^{-1}}\Phi(|\mathbf{t}|)|\mathbf{t}|d\mathbf{t}, \quad k=1,2,\ldots.$$

Then we see that

$$\begin{aligned}
P_{1,1}(\mathbf{x}) &= L\sum_{j=1}^{\infty}|(M^T)^{-Jj}\mathbf{x}|^2(\mu_0+\cdots+\mu_{j-1})^2 \\
&\le L\left[\sum_{k=0}^{\infty}\left(\sum_{j=k+1}^{\infty}|(M^T)^{-Jj}\mathbf{x}|^2\mu_k^2\right)^{1/2}\right]^2 \\
&\le C_2\Big(\int_{\mathbb{R}^s}\Phi(|\mathbf{t}|)d\mathbf{t}\Big)^2.
\end{aligned} \tag{2.7}$$

As for $P_{1,2}(x)$, we first observe that

$$P_{1,2}(\mathbf{x}) = L\sum_{j=1}^{\infty}\left(\sum_{k=j}^{\infty}\mu_k\right)^2 \le C_3\left[\sum_{k=1}^{\infty}\left(\sum_{j=1}^{k-1}\mu_k^2\right)^{1/2}\right]^2 \tag{2.8}$$

$$\le C_4\left[\sum_{k=1}^{\infty}\sqrt{k}\int_{|(M^T)^{-Jk}\mathbf{x}|^{-1}\le|\mathbf{t}|\le|(M^T)^{-J(k+1)}\mathbf{x}|^{-1}}\Phi(|\mathbf{t}|)d\mathbf{t}\right]^2.$$

By (2.1), we see that if $\mathbf{x}\in\Omega_J(M)$ and

$$|\mathbf{t}| \ge \max(2, |(M^T)^{-Jk}\mathbf{x}|^{-1}),$$

then

$$|\mathbf{t}| \ge |(M^T)^{-Jk}\mathbf{x}|^{-1} = \frac{1}{|\mathbf{x}|}\frac{|\mathbf{x}|}{|(M^T)^{-J}\mathbf{x}|}\frac{|(M^T)^{-J}\mathbf{x}|}{|(M^T)^{-J2}\mathbf{x}|}\cdots\frac{|(M^T)^{-J(k-1)}\mathbf{x}|}{|(M^T)^{-Jk}\mathbf{x}|}$$

$$\ge \frac{\theta^k}{|\mathbf{x}|} \ge \frac{\theta^k}{\min\limits_{x\in\Omega_J(M)}|\mathbf{x}|}.$$

Hence $\sqrt{\ln|\mathbf{t}|}\ge C_5\sqrt{k}$, and it follows from (2.8) that

$$P_{1,2}(\mathbf{x}) \le C_6\left[\int_{\mathbb{R}^s}\Phi(|\mathbf{t}|)\left(1+\sqrt{\ln^+|\mathbf{t}|}\right)d\mathbf{t}\right]^2. \tag{2.9}$$

Setting (2.7) and (2.8) into (2.6), we obtain

$$P_1(\mathbf{x}) \le C_7\left[\int_{\mathbb{R}^s}\Phi(|\mathbf{t}|)\left(1+\sqrt{\ln^+|\mathbf{t}|}\right)d\mathbf{t}\right]^2. \tag{2.10}$$

Next, let us estimate the sum

$$P_2(\mathbf{x}) := \sum_{\ell=1}^{L}\sum_{j=0}^{\infty}|\widehat{\psi}_\ell((M^T)^{Jj}\mathbf{x})|^2.$$

To do so, we divide $\mathbb{R}^s$ into cubes

$$Q_{\mathbf{m}} := \mathbf{m} + Q_0, \quad \mathbf{m}\in\mathbb{Z}^s, \tag{2.11}$$

with

$$Q_{\mathbf{0}} := \{\mathbf{x}\in\mathbb{R}^s\colon\ 0\le x_\alpha<1, \quad \alpha=1,\ldots,s\}$$

and set

$$\psi_\ell(\mathbf{m};\mathbf{x}) := \psi_\ell(\mathbf{x})\chi_{Q_{\mathbf{m}}}(\mathbf{x}).$$

Then by the triangle inequality, we have

$$P_2(\mathbf{x}) \le C_8\left[\sum_{\mathbf{m}\in\mathbb{Z}^s}\sum_{\ell=1}^{L}\left(\sum_{j=0}^{\infty}|\widehat{\psi}_\ell(\mathbf{m};(M^T)^{Jj}\mathbf{x})|^2\right)^{1/2}\right]^2. \tag{2.12}$$

We claim that there exists a constant C_9 depending only on s and M such that

$$\sum_{j=0}^{\infty}\left|\int_{Q_{\mathbf{m}}} g(\mathbf{t})e^{-i\mathbf{t}(M^T)^{Jj}\mathbf{x}}d\mathbf{t}\right|^2 \le C_9\int_{Q_{\mathbf{m}}}|g(\mathbf{t})|^2 d\mathbf{t}. \tag{2.13}$$

By Lemma 2, we see that this claim is equivalent to

$$\int_{Q_{\mathbf{m}}}\left|\sum_{j=0}^{\infty} c_j e^{-i\mathbf{t}(M^T)^{Jj}\mathbf{x}}\right|^2 d\mathbf{t} \le C_9\sum_{j=0}^{\infty}|c_j|^2, \quad \{c_j\}\in \ell^2. \tag{2.14}$$

It is sufficient to prove (2.14) for $\mathbf{m}=\mathbf{0}$. For this purpose, we decompose Q_0 into 2^s equal cubes and denote by $\mathbf{a}_\alpha$, $\alpha = 0, 1, \ldots, 2^s-1$, with $\mathbf{a}_0 = 0$ the vertices of these cubes with smallest componentwise coordinates. By the Bessel inequality for trigonometric systems, it is obvious that

$$\int_{Q_{\mathbf{0}}}\left|\sum_{\alpha=0}^{2^s-1}\sum_{\mathbf{k}\in\mathbb{Z}^s} b_{\alpha,\mathbf{k}} e^{-i\mathbf{t}(\mathbf{a}_\alpha+\mathbf{k})}\right|^2 dt \le C_{10}\sum_{\alpha,\mathbf{k}}|b_{\alpha,\mathbf{k}}|^2, \quad \text{all } \{b_{\alpha,\mathbf{k}}\}\in\ell^2, \tag{2.15}$$

where the constant C_{10} depends only on s. Now consider the cubes

$$R_{\alpha,\mathbf{k}} := \frac{1}{2}Q_0 + \mathbf{a}_\alpha + \mathbf{k}.$$

Since M satisfies (1.3), we see that there exist at most

$$\lambda := \left[\ln\sqrt{s}\Big/\ln\frac{1}{|||M^{-J}|||}\right]+1$$

numbers of points of the form $(M^T)^{Jj}\mathbf{x}$ in each cube $R_{\alpha,\mathbf{k}}$, where $j\ge 0$, $\mathbf{x}\in\Omega_J(M)$ and $|||\cdot|||$ denotes the spectral norm of M. Therefore, by (2.15), we obtain

$$\int_{Q_{\mathbf{0}}}\left|\sum_{\alpha,\mathbf{k}}\ \sum_{(M^T)^{Jj}\mathbf{x}\in R_{\alpha,\mathbf{k}}} c_j e^{-i\mathbf{t}(\mathbf{a}_\alpha+\mathbf{k})}\right|^2 d\mathbf{t} \le C_{11}\sum_j |c_j|^2, \quad \{c_j\}\in\ell^2,$$

where the constant C_{11} depends only on s and M. Thus to establish (2.14), it is sufficient to prove that for any $\{c_j\}\in\ell^2$, we have

$$\int_{Q_{\mathbf{0}}}\left|\sum_{\alpha,\mathbf{k}}\ \sum_{(M^T)^{Jj}\mathbf{x}\in R_{\alpha,\mathbf{k}}} c_j e^{-i\mathbf{t}(\mathbf{a}_\alpha+\mathbf{k})}(1-e^{-i\mathbf{t}\mathbf{h}(j,\alpha,\mathbf{k},\mathbf{x})})\right|^2 d\mathbf{t} \le C_{12}\sum|c_j|^2, \tag{2.16}$$

where

$$\mathbf{h}(j,\alpha,\mathbf{k},\mathbf{x}) = (h_{j,1},h_{j,2},\ldots,h_{j,s}) := (M^T)^{Jj}\mathbf{x} - (\mathbf{a}_\alpha+\mathbf{k})$$

with

$$|h_{j,\alpha}| \le \frac{1}{4}, \quad \alpha = 1,\ldots,s. \tag{2.17}$$

From (2.3), we see that

$$\begin{aligned}1-e^{-i\mathbf{t}\mathbf{h}(j,\alpha,\mathbf{k},\mathbf{x})} &= 1-\prod_{\beta=1}^{s}\{1-F_1(h_{j,\beta})-F_2(h_{j,\beta},t_\beta)-F_3(h_{j,\beta},t_\beta)\}\\ &= \sum{}' F_1(h_{j,\beta_1})\cdots F_1(h_{j,\beta_p})F_2(h_{j,\beta_{p+1}},t_{\beta_{p+1}})\cdots F_2(h_{j,\beta_q},t_{\beta_q})\\ &\quad\times F_3(h_{j,\beta_{q+1}},t_{\beta_{q+1}})\cdots F_3(h_{j,\beta_r},t_{\beta_r}),\end{aligned}\tag{2.18}$$

where $(\beta_1,\dots,\beta_r)$ is a permutation of some subset of the set $(1,\dots,s)$ and the sum $\sum'$ is taken over all possible nonempty subsets of $(1,\dots,s)$ and their permutations. According to (2.16), we need to estimate the integral

$$\begin{aligned}I := \int_{Q_\mathbf{0}}\Bigg|&\sum_{\alpha,k}\sum_{(M^T)^{Jj}\mathbf{x}\in R_{\alpha,k}} c_j e^{-i\mathbf{t}(\mathbf{a}_\alpha+\mathbf{k})}F_1(h_{j,\beta_1})\cdots F_1(h_{j,\beta_p})\\ &\times F_2(h_{j,\beta_{p+1}},t_{\beta_{p+1}})\cdots F_2(h_{j,\beta_q},t_{\beta_q})F_3(h_{j,\beta_{q+1}},t_{\beta_{q+1}})\\ &\cdots F_3(h_{j,\beta_r},t_{\beta_r})\Bigg|^2 d\mathbf{t}.\end{aligned}$$

By the generalized Minkowskii inequality and (2.17), we see that

$$\begin{aligned}I^{1/2} \le \sum_{n_{\beta_{p+1}}=1}^{\infty}\cdots\sum_{n_{\beta_r}=1}^{\infty}&\Bigg(\int_{Q_\mathbf{0}}\Bigg|\sum_{\alpha,\mathbf{k}}\sum_{(M^T)^{Jj}\mathbf{x}\in R_{\alpha,\mathbf{k}}} c_j e^{-i\mathbf{t}(\mathbf{a}_\alpha+\mathbf{k})}\\ &\times F_1(h_{j,\beta_1})\cdots F_1(h_{j,\beta_p})\frac{(-1)^{n_{\beta_{p+1}}}2h_{j,\beta_{p+1}}\sin\pi h_{j,\beta_{p+1}}}{\pi(n_{\beta_{p+1}}^2-h_{j,\beta_{p+1}}^2)}\\ &\cdots\frac{(-1)^{n_{\beta_q}}2h_{j,\beta_q}\sin\pi h_{j,\beta_q}}{\pi(n_{\beta_q}^2-h_{j,\beta_q}^2)}\,\frac{(-1)^{n_{\beta_{q+1}}}2h_{j,\beta_{q+1}}\cos\pi h_{j,\beta_{q+1}}}{\pi\left(\left(n_{\beta_{q+1}}-\frac12\right)^2-h_{j,\beta_{q+1}}^2\right)}\\ &\cdots\frac{(-1)^{n_{\beta_r}}2h_{j,\beta_r}\cos\pi h_{j,\beta_r}}{\pi\left(\left(n_{\beta_r}-\frac12\right)^2-h_{j,\beta_r}^2\right)}\cos n_{\beta_{p+1}}t_{\beta_{p+1}}\cdots\cos n_{\beta_q}t_{\beta_q}\\ &\times\sin\left(n_{\beta_{q+1}}-\frac12\right)t_{\beta_{q+1}}\cdots\sin\left(n_{\beta_r}-\frac12\right)t_{\beta_r}\Bigg|^2 d\mathbf{t}\Bigg)^{1/2}\\ \le C_{12}&\left(1-\frac{\sin\frac{\pi}{4}}{\frac{\pi}{4}}\right)^p\left(\sum_{n=1}^{\infty}\frac{1}{\pi\left(n^2-\frac{1}{16}\right)}\right)^{q-p}\\ &\times\left(\sum_{n=1}^{\infty}\frac{1}{\pi\left(\left(n-\frac12\right)^2-\frac{1}{16}\right)}\right)^{r-q}\left(\sum_j|c_j|^2\right)^{1/2}\\ \le C_{13}&\left(\sum|c_j|^2\right)^{1/2}.\end{aligned}\tag{2.19}$$

Hence by (2.18) and (2.19), we arrive at (2.14) and then (2.13). By (2.13), we then see that

$$\begin{aligned}\sum_{j=0}^{\infty}|\widehat{\psi}_\ell(\mathbf{m};(M^T)^{Jj}\mathbf{x})|^2 &= \sum_{j=0}^{\infty}\left|\int_{Q_\mathbf{m}}\psi_\ell(\mathbf{t})e^{-i\mathbf{t}(M^T)^{Jj}\mathbf{x}}d\mathbf{t}\right|^2\\ &\le C_{15}\int_{Q_\mathbf{m}}|\psi_\ell(\mathbf{t})|^2d\mathbf{t}.\end{aligned}$$

Therefore, it follows from (1.6) that

$$P_2(\mathbf{x}) \le C_{16} \left[\sum_{\mathbf{m}\in\mathbf{Z}^s} \left(\int_{Q_\mathbf{m}} (\Phi(|\mathbf{t}|))^2 d\mathbf{t} \right)^{1/2} \right]^2 \le C_{17} \left(\int_{\mathbb{R}^s} \Phi(|\mathbf{t}|) d\mathbf{t} \right)^2 . \tag{2.20}$$

Combining (2.10) and (2.20), we obtain (2.5). It is clear that (2.5) implies (1.12).

Conversely, by the hypothesis stated in Theorem 1, the functions $\widehat{\psi}_\ell$ are continuous on $\mathbb{R}^s$. Therefore, if $\widehat{\psi}_\ell(\mathbf{0}) \ne 0$ for some ℓ, then there exists a ball $B_\varepsilon := \{\mathbf{x}\colon |\mathbf{x}| \le \varepsilon\}$ on which $|\widehat{\psi}_\ell(\mathbf{x})|^2 \ge \eta > 0$. By (1.1), for any $\mathbf{x} \in \mathbb{R}^s \backslash \{\mathbf{0}\}$, there exists a j_0 such that $(M^T)^j \mathbf{x} \in B_\varepsilon$ for all $j \le j_0$. Therefore, $L_\Psi(\mathbf{x}) = \infty$. This is a contradiction to the hypothesis and hence completes the proof of Theorem 1(ii).

2.4. Counterexamples. Finally, we will construct examples of $\psi_\ell \in L^1 \cap L^2$, $\ell = 1, \ldots, L$, each of which has zero mean, but $\|L_\Psi\|^* = \infty$. For this purpose, it is sufficient to consider $s = 1$, $L = 1$, and $M = (2)$. Denote by h the hat function with supp $h = [-1, 1]$ defined by $h(x) := 1 - |x|$ for $|x| \le 1$. Set

$$\psi(x) := \sum_{n=0}^{\infty} 4^{-n} (h(x - m_n) + h(x + m_n)),$$

with $m_n = \prod_{k=1}^{n} 2^{4^{2k}}$. Then it is evident that $\psi \in L^1 \cap L^2$ and $\int \psi = 0$. We also observe that

$$\widehat{\psi}(x) = -\frac{8i \sin^2 \frac{x}{2}}{x^2} \sum_{n=0}^{\infty} 4^{-n} \sin m_n x. \tag{2.21}$$

Next, we prove that the function

$$L_\Psi(x) = \sum_{j\in\mathbf{Z}} |\widehat{\psi}(2^j x)|^2$$

is not bounded. To do so, consider

$$P_1(x) := \sum_{j=0}^{\infty} |\widehat{\psi}(2^{-j} x)|^2 .$$

Then since $\|L_\Psi\|^* \ge \|P_1\|_{L[0,1]}$, it is sufficient to show that

$$\|P_1\|_{L[0,1]} = \infty. \tag{2.22}$$

To prove (2.22), we see that

$$\begin{aligned}
\int_0^1 |P_1(x)|dx &= C_{18}\sum_{j=0}^{\infty} 2^{3j}\int_{2^{-j}}^{2^{-j+1}} \frac{1}{2^{2j}}\left|\sum_{n=1}^{\infty} 4^{-n}\sin m_n x\right|^2 dx \\
&\geq C_{19}\int_0^1 \frac{1}{x}\left|\sum_{n=1}^{\infty} 4^{-n}\sin m_n x\right|^2 dx - O(1) \\
&\geq C_{19}\sum_{k=1}^{\infty}\int_{m_{k+1}^{-1}}^{m_k^{-1}} \frac{1}{x}\left|\sum_{n=k+1}^{\infty} 4^{-n}\sin m_n x\right|^2 dx - O(1) \\
&= C_{19}\sum_{k=1}^{\infty}\int_{1}^{m_{k+1}m_k^{-1}} \frac{1}{x}\left(\sum_{n=k+1}^{\infty} 4^{-n}\sin \frac{m_n}{m_{k+1}}x\right)^2 dx - O(1) \\
&\geq C_{19}\sum_{k=1}^{\infty}\sum_{\nu=1}^{\left[\frac{m_{k+1}}{2\pi m_k}\right]-1} \frac{1}{2(\nu+1)\pi}\int_{2\nu n}^{2(\nu+1)\pi}\left|\sum_{n=k+1}^{\infty} 4^{-n}\sin \frac{m_n}{m_{k+1}}x\right|^2 dx - O(1) \\
&= C_{19}\sum_{k=1}^{\infty}\sum_{\nu=1}^{\left[\frac{m_{k+1}}{2\pi m_k}\right]-1} \frac{1}{\nu+1}\sum_{n=k+1}^{\infty} 4^{-2n} - O(1) = \infty.
\end{aligned}$$

This completes the proof of Theorem 1. □

3. Proof of Theorem 2. To prove Theorem 2, we consider the sum

$$\begin{aligned}
I_{J'} := \sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'} \frac{|\det M|^j}{(2\pi)^s}&\int_{[(M^T)^{-j}\mathbf{x}_0-\pi b^{-1}\mathbf{e},(M^T)^{-j}\mathbf{x}_0+\pi b^{-1}\mathbf{e}]} \\
&\left|\sum_{\mathbf{m}\in\mathbb{Z}^s} \hat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi_\ell}(\mathbf{x}+2\pi b^{-1}\mathbf{m})}\right|^2 d\mathbf{x},
\end{aligned}$$

where $\widehat{f}(\mathbf{x}) := (\pi\varepsilon^{-1})^{s/2}\chi_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}(\mathbf{x})$ as in the proof of Theorem 1(i). For any fixed $j \in \mathbb{Z}$, we decompose $\mathbb{Z}^s$ into a union of two disjoint sets

$$\mathbb{Z}^s = \mathbb{Z}_{j,1} \cup \mathbb{Z}_{j,2},$$

where $\mathbb{Z}_{j,2}$ is the set of all $\mathbf{m}$ for which the function $\hat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))$ is equal to zero everywhere on the cube $[(M^T)^{-j}\mathbf{x}_0 - \pi b^{-1}\mathbf{e},\ (M^T)^{-j}\mathbf{x}_0 + \pi b^{-1}\mathbf{e}]$. If $|(M^T)^j\mathbf{e}| > \pi^{-1}b\varepsilon$, then the cardinality $n(\mathbb{Z}_{j,1})$ of the set $\mathbb{Z}_{j,1}$ is bounded. If $|(M^T)^j\mathbf{e}| \leq \pi^{-1}b\varepsilon$, then $n(\mathbb{Z}_{j,1}) = O(\varepsilon^s|\det M|^j)$. Thus we have

$$n(\mathbb{Z}_{j,1}) = O(\varepsilon^s|\det M|^j + 1). \tag{3.1}$$

It follows from (3.2) that the function

$$g_j(x) := \sum_{\mathbf{m}\in\mathbb{Z}^s} \chi_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))$$

is dominated by $O(\varepsilon^s|\det M|^j+1)$, where "$O$" is independent of ε and j. Hence by the Cauchy inequality, we see that

$$\begin{aligned} I_{J'} &\le \sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'}\frac{|\det M|^j}{(2\varepsilon)^s}\int_{[(M^T)^{-j}\mathbf{x}_0-\pi b^{-1}\mathbf{e},(M^T)^{-j}\mathbf{x}_0+\pi b^{-1}\mathbf{e}]} g_j(\mathbf{x}) \\ &\quad\times\sum_{\mathbf{m}\in\mathbf{Z}^s}|\widehat{\psi}_\ell(\mathbf{x}+2\pi\mathbf{m})|^2\chi_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))d\mathbf{x} \qquad (3.2)\\ &\le C_1\sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'}\int_{(M^T)^{-j}([\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}])}\left\{|\widehat{\psi}_\ell(\mathbf{x})|^2+\frac{|\det M|^j}{(2\varepsilon)^s}|\widehat{\psi}_\ell(\mathbf{x})|^2\right\}d\mathbf{x}. \end{aligned}$$

Since $|||M^{-1}|||<\infty$, we may choose for any fixed $\mathbf{x}_0\neq\mathbf{0}$ and any ball $B_r:=\{\mathbf{x}\colon|\mathbf{x}|\le r\}$ some $\varepsilon_0>0$ and sufficiently large J_0 such that for any $0<\varepsilon<\varepsilon_0$ and $J'>J_0$, the sets $(M^T)^{-j}([\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}])$, $-\infty<j\le -J'$, are mutually disjoint and are also disjoint with B_r. Hence for an arbitrarily given $\eta>0$, we have

$$\sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'}\int_{(M^T)^{-j}([\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}])}|\widehat{\psi}_\ell(\mathbf{x})|^2d\mathbf{x}\le\eta$$

for all sufficiently large J and sufficiently small $\varepsilon>0$. Therefore, by (3.2), we get

$$\begin{aligned} I'_{J'} &:= \sum_{\ell=1}^{L}\sum_{j=-J'+1}^{\infty}\frac{|\det M|^j}{(2\pi)^s}\int_{[0,2\pi b^{-1}]^s} \qquad (3.3)\\ &\quad\left|\sum_{\mathbf{m}\in\mathbf{Z}^s}\hat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi}_\ell(\mathbf{x}+2\pi b^{-1}\mathbf{m})}\right|^2d\mathbf{x} \\ &= \langle T_bf,f\rangle - I_{J'}\ge\|T_b\|_*-C_2\eta \\ &\quad-\frac{C_3}{(2\varepsilon)^s}\int_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}\sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'}|\widehat{\psi}_\ell((M^T)^{-j}\mathbf{x})|^2d\mathbf{x}. \end{aligned}$$

On the other hand, it is clear that for any fixed J and all sufficiently small $\varepsilon>0$, we have

$$\begin{aligned} I'_{J'} &= \sum_{\ell=1}^{L}\sum_{j=-J'+1}^{\infty}|\det M|^j\int_{(M^T)^{-j}([\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}])}|\hat{f}((M^T)^j\mathbf{x})\overline{\widehat{\psi}_\ell(\mathbf{x})}|^2d\mathbf{x} \\ &= \sum_{\ell=1}^{L}\frac{1}{(2\varepsilon)^s}\int_{[\mathbf{x}_0-\varepsilon\mathbf{e},\mathbf{x}_0+\varepsilon\mathbf{e}]}\sum_{j=-J'+1}^{\infty}|\widehat{\psi}_\ell((M^T)^{-j}\mathbf{x})|^2d\mathbf{x}. \end{aligned}$$

Hence in view of the boundedness of L_Ψ, we may take $\varepsilon\to0$ in (3.3) to arrive at

$$\sum_{\ell=1}^{L}\sum_{j=-J'+1}^{\infty}|\widehat{\psi}_\ell((M^T)^{-j}\mathbf{x}_0)|^2\ge\|T_b\|_*-C_2\eta-C_3\sum_{\ell=1}^{L}\sum_{j=-\infty}^{-J'}|\widehat{\psi}_\ell((M^T)^{-j}\mathbf{x}_0)|^2 \qquad (3.4)$$

for almost all x_0. Since η is arbitrary, assertion (3.4) together with the boundedness of L_Ψ yields $\|T_b\|_*\le\|L_\Psi\|_*$. This completes the proof of Theorem 2. □

In the following, we consider the special case where $\operatorname{supp} \widehat{\psi}_\ell \subset [-\pi/b, \pi/b]^s$, $\ell = 1, 2, \dots, L$. For this case, we have

$$\begin{aligned}\langle T_b f, f\rangle &= \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \frac{|\det M|^j}{(2\pi)^s} \int_{[0,2\pi b^{-1}]^s} \left| \sum_{\mathbf{m}\in\mathbb{Z}^s} \hat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\overline{\widehat{\psi}_\ell(\mathbf{x}+2\pi b^{-1}\mathbf{m})} \right|^2 d\mathbf{x} \\ &= \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \frac{|\det M|^j}{(2\pi)^s} \int_{[0,2\pi b^{-1}]^s} \sum_{\mathbf{m}\in\mathbb{Z}^s} |\hat{f}((M^T)^j(\mathbf{x}+2\pi b^{-1}\mathbf{m}))\widehat{\psi}_\ell(\mathbf{x}+2\pi b^{-1}\mathbf{m})|^s d\mathbf{x} \\ &= \sum_{\ell=1}^{L} \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} \sum_{j\in\mathbb{Z}} |\widehat{\psi}_\ell((M^T)^{-j}\mathbf{x})|^2 |\hat{f}(\mathbf{x})|^2 d\mathbf{x}.\end{aligned}$$

This implies that

$$\|T_b\|_* = \|L_\Psi\|_* \quad \text{and} \quad \|T_b\|^* = \|L_\Psi\|^*. \tag{3.5}$$

A more general formulation can be stated as follows. Set

$$E(\psi) := \{\omega\colon\ \widehat{\psi}(\omega) = 0\},$$

and for $\mathbf{k} \in \mathbb{Z}^s$, $\ell = 1, \dots, L$, consider

$$E_{\mathbf{k}}(\psi_\ell) := \{\omega \in [-\pi/b, \pi/b]^s\colon\ \omega + 2\mathbf{k}\pi/b \in E(\psi_\ell)\}.$$

Then we have the following.

Corollary 1. *Let $\psi_\ell \in L^2$, $\ell = 1, \dots, L$, such that for each ℓ, the sets $E_{\mathbf{k}}(\psi_\ell)$, $\mathbf{k} \in \mathbb{Z}^s$, are mutually disjoint. Then* (3.5) *holds.*

4. Proof of Theorem 3.

In this section, we establish a general result which implies the multivariate (matrix-dilated) version of the so-called second oversampling theorem introduced in [3].

4.1. A preliminary result.

In order to prove Theorem 3, we first introduce some notations and establish a lemma. For a natural number $n \geq 2$, consider

$$\mathcal{B}_n := \{\mathbf{p} = (p_1, \dots, p_s)\colon\ p_\ell \in \mathbb{Z}, \quad 0 \leq p_\ell \leq n-1, \quad \ell = 1, \dots, s\}$$

and adopt the notation

$$\mathcal{B}_n^0 := \mathcal{B}_n \backslash \{\mathbf{0}\}.$$

Let $\mathbf{a} = (a_1, \dots, a_s)$ and $\mathbf{b} = (b_1, \dots, b_s)$ be two integer vectors. If $\mathbf{a}$ and $\mathbf{b}$ satisfy $a_\ell \equiv b_\ell \pmod{n}$, $\ell = 1, \dots, s$, then we write

$$\mathbf{a} \equiv \mathbf{b} \pmod{n}.$$

We have the following result.

Lemma 4. *Let M be an integer matrix with $|\det M| > 1$ and $n \geq 2$ be any natural number that satisfies* (1.15). *Then the equation*

$$M\mathbf{p} \equiv \mathbf{0} \pmod{n} \tag{4.1}$$

does not have any solution in $\mathcal{B}_n^0$.

Proof. Suppose that $\mathbf{p} = (p_1, \ldots, p_s) \in \mathcal{B}_n^0$ satisfies (4.1). Then there exists some $\mathbf{k} = (k_1, \ldots, k_s) \in \mathbb{Z}^s$ that satisfies

$$M\mathbf{p} = n\mathbf{k}.$$

Therefore, we have for some ℓ, $1 \le \ell \le s$, that

$$1 \le p_\ell \le n - 1 \tag{4.2}$$

and

$$p_\ell = \frac{n\Delta_\ell}{\det M}, \tag{4.3}$$

where

$$\Delta_\ell = \begin{vmatrix} m_{11} & \ldots & m_{1,\ell-1} & k_1 & m_{1,\ell+1} & \ldots & m_{1s} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ m_{s1} & \ldots & m_{s,\ell-1} & k_s & m_{s,\ell+1} & \ldots & m_{ss} \end{vmatrix}.$$

Since $(n, |\det M|) = 1$ and p_ℓ is an integer, it follows from (4.3) that $\Delta_\ell / \det M$ must also be an integer. Hence (4.3) implies that p_ℓ / n is an integer as well. However, this is a contradiction to (4.2). □

4.2. Proof of Theorem 3. The proof of this theorem is divided into three steps.

4.2.1. Decomposition. Let $\mathbf{p}^1$ be any vector in $\mathcal{B}_n^0$ and consider the sequence of vectors

$$M^0\mathbf{p}^1 = \mathbf{p}^1, \quad M\mathbf{p}^1, \quad M^2\mathbf{p}^1, \ldots.$$

By Lemma 4, we see that to every vector $M^j\mathbf{p}^1$ in this sequence, there exists a vector $\mathbf{q}_j = (q_1^j, \ldots, q_s^j) \in \mathcal{B}_n^0$ that satisfies

$$M^j\mathbf{p}^1 \equiv \mathbf{q}_j \pmod{n}.$$

Since $\mathcal{B}_n^0$ is a finite set, there exist $0 \le j_1 < j_2$ such that

$$\mathbf{q}_{j_1} = \mathbf{q}_{j_2}. \tag{4.4}$$

If j_1 and j_2 are the smallest integers for which (4.4) holds, then j_1 must be zero. Therefore, we obtain an ordered system $\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{j_2}$ with $j_2 > 0$, $\mathbf{q}_0 = \mathbf{p}^1 = \mathbf{q}_{j_2}$, and $\mathbf{q}_j \ne \mathbf{q}_0$, where $0 < j < j_2$. Denote this system by $\mathcal{F}_1 = (\mathbf{p}_1^1, \ldots, \mathbf{p}_{\mu_1}^1)$, i.e., $\mathbf{p}_1^1 = \mathbf{q}_0, \mathbf{p}_2^1 = \mathbf{q}_1, \ldots, \mathbf{p}_{\mu_1}^1 = \mathbf{q}_{j_2}$. If $\mathcal{F}_1 = \mathcal{B}_n^0$, we terminate the process. Otherwise, we choose $\mathbf{p}_1^2 \in \mathcal{B}_n^0 \backslash \mathcal{F}_1$ and the same process yields the second ordered system $\mathcal{F}_2 = (\mathbf{p}_1^2, \ldots, \mathbf{p}_{\mu_2}^2)$. Continue this process until all the elements in $\mathcal{B}_n^0$ are chosen. Then we arrive at a decomposition

$$\mathcal{B}_n^0 = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_r,$$

which has the following two properties:

(a) $\mathcal{F}_t \cap \mathcal{F}_{t'} = \emptyset$ if $t \ne t'$.

(b) The periodic extension $\mathcal{F}_t^p$ of $\mathcal{F}_t = (\mathbf{p}_1^t, \ldots, \mathbf{p}_{\mu_t}^t)$, defined by

$$\mathcal{F}_t^p = (\ldots \mathbf{p}_{-1}^t, \mathbf{p}_0^t, \mathbf{p}_1^t, \ldots, \mathbf{p}_{\mu_t}^t, \ldots) \quad \text{and} \quad \mathbf{p}_{\mu_t^+\nu}^t = \mathbf{p}_\nu^t \quad \text{for all } \nu \in \mathbb{Z},$$

satisfies $M\mathbf{p}_j^t \equiv \mathbf{p}_{j+1}^t (\mathrm{mod}\ n)$, where $t = 1, \dots, r$, $j \in \mathbb{Z}$.

We now decompose the family of functions

$$\mathcal{S}_\ell := \{\psi_{\ell,b/n;j,\mathbf{k}}(\mathbf{x})\colon\ 1 \le \ell \le L, \quad j \in \mathbb{Z}, \quad \mathbf{k} \in \mathbb{Z}^s\}$$

into n^s subsets as follows:

$$\mathcal{S}_\ell = \mathcal{S}_{\ell,0} \cup \bigcup_{t=1}^{r} \{\mathcal{S}_{\ell,1}^t \cup \cdots \cup \mathcal{S}_{\ell,\mu_t}^t\},$$

where

$$\mathcal{S}_{\ell,0} := \{\psi_{\ell,b;j,\mathbf{k}}(\mathbf{x})\colon\ 1 \le \ell \le L, \quad j \in \mathbb{Z}, \quad \mathbf{k} \in \mathbb{Z}^s\}$$

and

$$\mathcal{S}_{\ell,\mu}^t := \left\{ |\det M|^{j/2} \psi_\ell \left(M^j \mathbf{x} - \frac{b\mathbf{p}_{j+\mu}^t}{n} - \mathbf{k}b \right)\colon\ 1 \le \ell \le L, \quad j \in \mathbb{Z}, \quad \mathbf{k} \in \mathbb{Z}^s \right\},$$

where $\mu = 1, \dots, \mu_\ell$ and $t = 1, \dots, r$. By the definition of $\|T_b\|_*$ and $\|T_b\|^*$, we see that for any f with $\|f\| = 1$, we have

$$\|T_b\|_* \le \sum_{\ell=1}^{L} \sum_{g \in \mathcal{S}_{\ell,0}} |\langle f, g \rangle|^2 \le \|T_b\|^*. \tag{4.5}$$

4.2.2. Upper bound. Let $j_0 \in \mathbb{Z}$ and consider

$$\sigma_{\ell,j_0}^{\mu,t}(f) := \sum_{j \ge j_0} \sum_{\mathbf{k}} \left| \left\langle |\det M|^{j/2} \psi_\ell \left(M^j \cdot - \frac{\mathbf{p}_{j+\mu}^t b}{n} - \mathbf{k}b \right), f \right\rangle \right|^2,$$

where $\mu = 1, \dots, \mu_t$ and $t = 1, \dots, r$. If $j \ge j_0$, then since the property (b) implies that

$$M^{j-j_0} \mathbf{p}_{j_0+\mu}^t \equiv \mathbf{p}_{j+\mu}^t \quad (\mathrm{mod}\ n),$$

for any $\mathbf{k} \in \mathbb{Z}^s$, we have a unique $\mathbf{k'} \in \mathbb{Z}^s$ such that

$$M^j \mathbf{x} - \frac{b\mathbf{p}_{j+\mu}^t}{n} - \mathbf{k}b = M^j \left(\mathbf{x} - \frac{bM^{-j_0} \mathbf{p}_{j_0+\mu}^t}{n} \right) - \mathbf{k'}b.$$

Therefore, the two collections of functions $\{M^j \mathbf{x} - b\mathbf{p}_{j+\mu}^t/n - \mathbf{k}b\colon \mathbf{k} \in \mathbb{Z}^s\}$ and $\{M^j(\mathbf{x} - bM^{-j_0}\mathbf{p}_{j_0+\mu}^t/n) - \mathbf{k}b\colon\ \mathbf{k} \in \mathbb{Z}^s\}$ are identical. It then follows for any $f \in L^2$ with $\|f\| = 1$ that

$$\begin{aligned}
\sum_{\ell=1}^{L} \sigma_{\ell,j_0}^{\mu,t}(f) &= \sum_{\ell=1}^{L} \sum_{j \ge j_0} \sum_{k \in \mathbb{Z}^s} \left| \left\langle |\det M|^{j/2} \psi_\ell \left(M^j \left(\cdot - \frac{bM^{-j_0}\mathbf{p}_{j_0+\mu}^t}{n} \right) - \mathbf{k}b \right), f \right\rangle \right|^2 \\
&= \sum_{\ell=1}^{L} \sum_{j \ge j_0} \sum_{\mathbf{k} \in \mathbb{Z}^s} \left| \left\langle |\det M|^{j/2} \psi_\ell (M^j \cdot - \mathbf{k}), f \left(\cdot + \frac{bM^{-j_0}\mathbf{p}_{j_0+\mu}^t}{n} \right) \right\rangle \right|^2 \\
&\le \|T_b\|^*.
\end{aligned}$$

Thus we obtain

$$\sum_{\ell=1}^{L} \sum_{g\in\mathcal{S}^t_{\ell,\mu}} |\langle f, g\rangle|^2 = \lim_{j_0\to-\infty} \sum_{\ell=1}^{L} \sigma^{\mu,t}_{\ell,j_0}(f) \le \|T_b\|^*. \tag{4.6}$$

Combining (4.5) and (4.6), we arrive at

$$\begin{aligned}
&\sum_{\ell=1}^{L} \sum_{j\in\mathbf{Z}} \sum_{\mathbf{k}\in\mathbf{Z}^s} |\langle f, \psi_{b/n,\ell;j,\mathbf{k}}\rangle|^2 \\
&\qquad = \sum_{\ell=1}^{L} \sum_{g\in\mathcal{S}_{\ell,0}} |\langle f, g\rangle|^2 + \sum_{\ell=1}^{L} \sum_{t=1}^{\mu} \sum_{\mu=1}^{\mu_t} \sum_{g\in\mathcal{S}^t_{\ell,\mu}} |\langle f, g\rangle|^2 \\
&\qquad \le n^s \|T_b\|^*
\end{aligned}$$

for any $f \in L^2$ with $\|f\| = 1$. Therefore, we have $\|T_{b/n}\|^* \le \|T_b\|^*$.

4.2.3. Lower bound. To establish a lower-bound estimate for $\|T_{b/n}\|_*$, it is sufficient to restrict our attention to compactly supported and bounded functions f with $\|f\| = 1$. Suppose that supp $f \subset [-K, K]^s$, $K > 0$. Set

$$\theta^{\mu,t}_{\ell,j_0}(f) := \sum_{j<j_0} \sum_{\mathbf{k}\in\mathbf{Z}^s} |\langle |\det M|^{j/2} \psi_\ell(M^j \cdot -\mathbf{k}b), f^{\mu,t}_{j_0}\rangle|^2,$$

where

$$f^{\mu,t}_{j_0}(\mathbf{x}) := f\left(\mathbf{x} + \frac{bM^{-j_0}\mathbf{p}^t_{j_0+\mu}}{n}\right). \tag{4.7}$$

Then we have

$$\begin{aligned}
\sum_{\ell=1}^{L} \sum_{g\in\mathcal{S}^t_{\ell,\mu}} |\langle f, g\rangle|^2 &= \sum_{\ell=1}^{L} \sum_{j\in\mathbf{Z}} \sum_{\mathbf{k}\in\mathbf{Z}^s} \left|\left\langle |\det M|^{j/2} \psi_\ell\left(M^j \cdot - \frac{\mathbf{p}^t_{j+\mu} b}{n} - \mathbf{k}b\right), f\right\rangle\right|^2 \\
&\ge \sum_{\ell=1}^{L} \sigma^{\mu,t}_{\ell,j_0}(f) + \sum_{\ell=1}^{L} \theta^{\mu,t}_{\ell,j_0}(f) - \sum_{\ell=1}^{L} \theta^{\mu,t}_{\ell,j_0}(f) \qquad (4.8)\\
&\ge \|T\|_* - \sum_{\ell=1}^{L} \theta^{\mu,t}_{\ell,j_0}(f).
\end{aligned}$$

We claim that

$$\lim_{j_0\to-\infty} \theta^{\mu,t}_{\ell,j_0}(f) = 0. \tag{4.9}$$

To prove (4.9), we first observe that by (4.7), supp $f^{\mu,t}_0 \subset I^{\mu,t}_{j_0}$, where

$$I^{\mu,t}_{j_0} := \left[-K\mathbf{e} - \frac{bM^{-j_0}\mathbf{p}^t_{j_0+\mu}}{n}, K\mathbf{e} - \frac{bM^{-j_0}\mathbf{p}^t_{j_0+\mu}}{n}\right].$$

Hence by the Cauchy inequality, we see that

$$\begin{aligned}|\langle|\det M|^{j/2}\psi_\ell(M^j\cdot-\mathbf{k}),f_{j_0}^{\mu,t}\rangle|^2 &\le |\det M|^j(2K)^s\|f\|_\infty\int_{I_{j_0}^{\mu,t}}|\psi_\ell(M^j\mathbf{x}-\mathbf{k}b)|^2d\mathbf{x}\\ &=(2K)^s\|f\|_\infty\int_{M^j\left(I_{j_0}^{\mu,t}\right)}|\psi_\ell(\mathbf{x}-\mathbf{k}b)|^2d\mathbf{x}.\end{aligned}$$

Thus for any positive J_0, we have

$$\begin{aligned}\theta_{\ell,j_0}^{\mu,t}(f)\le{}&(2K)^s\|f\|_\infty\sum_{j_0-J_0\le j<j_0}\sum_{\mathbf{k}\in\mathbf{Z}^s}\int_{\mathbf{k}b+M^j\left(I_{j_0}^{\mu,t}\right)}|\psi_\ell(\mathbf{x})|^2d\mathbf{x}\\ &+(2K)^s\|f\|_\infty\sum_{j\le j_0-J_0}\sum_{\mathbf{k}\in\mathbf{Z}^s}\int_{\mathbf{k}b+M^j\left(I_{j_0}^{\mu,t}\right)}|\psi_\ell(\mathbf{x})|^2d\mathbf{x}.\end{aligned}$$

On the other hand, observe that for $\mathbf{k}\neq\mathbf{k}'$ and sufficiently small j and j', we have

$$\{\mathbf{k}b+M^j(I_{j_0}^{\mu,t})\}\cap\{\mathbf{k}'b+M^{j'}(I_{j_0}^{\mu,t})\}=\emptyset.$$

If $\mathbf{k}$ is fixed, then the function

$$\sum_{j\le j_0-J_0}\chi_{\mathbf{k}b+M^j\left(I_{j_0}^{\mu,t}\right)}(\mathbf{x})$$

is bounded and

$$\lim_{J_0\to\infty}\left|\bigcup_{j\le j_0-J_0}(\mathbf{k}b+M^j(I_{j_0}^{\mu,t}))\right|=0.$$

Therefore, for any $\varepsilon>0$, since $\psi_\ell\in L^2$, we may find some positive J_0 such that for all $J'\ge J_0$,

$$\theta_{\ell,j_0}^{\mu,t}(f)\le C_1(2K)^s\|f\|_\infty\sum_{j_0-J'\le j<j_0}\sum_{\mathbf{k}\in\mathbf{Z}^s}\int_{\mathbf{k}b+M^j\left(I_{j_0}^{\mu,t}\right)}|\psi_\ell(\mathbf{x})|^2d\mathbf{x}+\varepsilon. \tag{4.10}$$

Fix $J'\ge J_0$. Then taking the limit of the quantity on the right-hand side of (4.10) as $j_0\to-\infty$, we get

$$\overline{\lim_{j_0\to\infty}}\,\theta_{\ell,j_0}^{\mu,t}(f)\le\varepsilon.$$

This establishes the claim in (4.9). Now applying (4.5), (4.8), and (4.9), we obtain

$$\begin{aligned}&\sum_{\ell=1}^{L}\sum_{j\in\mathbf{Z}}\sum_{\mathbf{k}\in\mathbf{Z}^s}|\langle f,\psi_{\ell,b/n;j,\mathbf{k}}\rangle|^2\\ &\quad=\sum_{\ell=1}^{L}\sum_{g\in\mathcal{S}_{\ell,0}}|\langle f,g\rangle|^2+\sum_{\ell=1}^{L}\sum_{t=1}^{r}\sum_{\mu=1}^{\mu_t}\sum_{g\in\mathcal{S}_{\ell,\mu}^t}|\langle f,g\rangle|^2\ge n^s\|T_b\|_*.\end{aligned}$$

That is, we have

$$\|T_{b/n}\|_*\ge\|T_b\|_*.$$

This completes the proof of (1.16).

That (1.16) does not hold in general without the assumption (1.15) follows from Remark 2 in [3; p. 45]. □

5. Applications to wavelet analysis and a study of the boundedness of g-function operators. In this section, we give some applications of Theorems 1–3 to the analysis of wavelets and affine frames and study the boundedness of the Littlewood–Paley g-function operators. In particular, a characterization of stable matrix-dilated wavelet families is given, and the second oversampling theorem for the matrix-dilated setting is established.

5.1. Matrix-dilated wavelets. Let M be an $s \times s$ real matrix that satisfies (1.3). A family $\Psi = \{\psi_1, \dots, \psi_L\} \subset L^2$ is called a *stable matrix-dilated wavelet family relative to* M if there exist two positive numbers A and B such that

$$A\|f\|^2 \leq \sum_{\ell=1}^{L} \sum_{j} \|W_{\ell,j} f\|^2 \leq B\|f\|^2, \quad f \in L^2, \tag{5.1}$$

where $W_{\ell,j} f$ is defined by

$$(W_{\ell,j} f)(\mathbf{x}) := f * (|\det M|^j \psi_\ell(M^j \cdot))(\mathbf{x}). \tag{5.2}$$

We call the constants A and B wavelet *bounds*.

Denote by $\eta_\ell(\mathbf{x})$, $\ell = 1, \dots, L$, the functions determined by

$$\hat{\eta}_\ell(\mathbf{x}) := \frac{\widehat{\psi}_\ell(\mathbf{x})}{\sum\limits_{p=1}^{L} \sum\limits_{j \in \mathbf{Z}} |\widehat{\psi}_p((M^T)^j \mathbf{x})|^2}. \tag{5.3}$$

If Ψ is a stable matrix-dilated wavelet family relative to M, then any $f \in L^2$ has a decomposition

$$f(x) = \sum_{\ell=1}^{L} \sum_{j \in \mathbf{Z}} ((W_{\ell,j} f) * \bar{\eta}_{\ell,j}^{-})(\mathbf{x}),$$

where $(W_{\ell,j} f)(\mathbf{x})$ is the integral wavelet transform of f relative to Ψ as defined in (5.2), $\eta_{\ell,j}(\mathbf{x}) := |\det M|^j \eta_\ell(M^j \mathbf{x})$ with η_ℓ given by (5.3), and $g^-(\mathbf{x})$ denotes $g(-\mathbf{x})$. For the one- and two-dimensional settings with $M = (2)$ and $M = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, respectively, this decomposition has been studied in the framework of signal analysis by Mallat and Zhong [5] (see also [4] and [6]). It is clear that (5.1) holds if and only if the corresponding Littlewood–Paley energy function (1.6) satisfies

$$A \leq L_\Psi(x) \leq B \quad \text{a.e.}$$

Therefore, using Theorem 1, we may establish the following characterization of stable matrix-dilated wavelet families.

PROPOSITION 1. *Let M be an $s \times s$ real matrix that satisfies* (1.3), *and assume that the family $\Psi = (\psi_1, \dots, \psi_L)$ satisfies* (1.11). *Then Ψ is a stable matrix-dilated wavelet family relative to M in the sense of* (5.1) *if and only if Ψ satisfies both of the following.*

(i) $\int \psi_\ell(\mathbf{x}) d\mathbf{x} = 0, \ell = 1, \dots, L$; *and*

(ii) $L_\Psi(\mathbf{x}) > 0$ *for any* $\mathbf{x} \in \Omega_J(M)$, *where* $\Omega_J(M)$ *is defined in* (2.2).

Indeed, by Theorem 1, we see that (i) is a necessary condition for Ψ to be a stable matrix-dilated wavelet family. It is also evident that (ii) is also a necessary condition. On the other hand, if Ψ satisfies (1.11), then by the proof of Theorem 1, we see that the series $\sum_{j\in\mathbb{Z}} |\widehat{\psi}_\ell((M^T)^j\mathbf{x})|^2$ is uniformly convergent to a continuous function on $\overline{\Omega_J(M)}$. Thus (i) and (ii) imply that Ψ is a stable matrix-dilated wavelet family.

As a consequence of this proposition, we have the following criterion on stable matrix-dilated wavelet families (see [2] for the one-variable setting).

Corollary 2. *Let M be an $s \times s$ real matrix that satisfies* (1.3) *and $\Psi = (\psi_1, \dots, \psi_L)$ be a family of functions that satisfy* (1.11) *and* (1.13). *If there exists some $j_0 \in \mathbb{Z}$ such that $\widehat{\psi}(\omega) \neq 0$ on the set*

$$(M^T)^{j_0 J}\overline{\Omega_J(M)} := \{\omega\colon\ (M^T)^{-j_0 J}\omega \in \overline{\Omega_J(M)}\},$$

where $\Omega_J(M)$ is defined in (2.2), *then Ψ is a stable matrix-dilated wavelet family relative to M.*

Recall that a family (1.2) is called a *frame* with frame bounds A and B, $0 < A \leq B$, if

$$A\|f\|^2 \leq \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z},\mathbf{k}\in\mathbb{Z}^s} |\langle f, \psi_{\ell,b;j,\mathbf{k}}\rangle|^2 \leq B\|f\|^2, \quad f \in L^2. \tag{5.4}$$

If $A = B$ in (5.4), then the family (1.2) is called a *tight frame.*

By Theorems 1 and 2, we may establish the following.

Proposition 2. *Let $M = \lambda U$, where $\lambda > 1$ and U is an $s \times s$ unitary matrix. If the family* (1.2) *constitutes a frame with frame bounds A and B, then Ψ is a stable matrix-dilated wavelet family relative to M with wavelet frame bounds A and B.*

Conversely, if ψ_ℓ, $\ell = 1, \dots, L$, satisfy the condition in Corollary 1 *and Ψ is a matrix-dilated wavelet family relative to M with bounds A and B, then* (1.2) *constitutes a frame with bounds A and B.*

For the univariate setting, the first statement of this result was established in our earlier work [1].

5.2. Oversampling for frames. As a consequence Theorem 3, we have the following so-called second oversampling theorem for frames.

Proposition 3. *Let all the entries of M be integers and the condition* (1.3) *be satisfied. If the family* (1.2) *constitutes a frame with frame bounds A and B and n is a natural number that satisfies* (1.15), *then the family*

$$\{n^{-s/2}\psi_{\ell,b/n;j,\mathbf{k}}(\mathbf{x})\colon\ 1 \leq \ell \leq L, \quad j \in \mathbb{Z}, \quad \mathbf{k} \in \mathbb{Z}^s\} \tag{5.5}$$

also constitutes a frame with the same bounds. However, this conclusion does not hold in general without assumption (1.13).

This is a generalization of Theorem 4 in [3]. From this proposition, we see that if M and n satisfy the assumptions in Proposition 3 and the family (1.2) constitutes a tight frame, then the family (5.5) is a tight frame as well.

5.3. Boundedness of g-function operators. Let ψ be any square-integrable function on $\mathbb{R}^s$ with $\int \psi = 0$ and set $\psi_t(\mathbf{x}) := t^{-s}\psi(\mathbf{x}/t)$. Then the operator

$$g(f)(\mathbf{x}) := \left(\int_0^\infty |\psi_t * f(\mathbf{x})|^2 \frac{dt}{t}\right)^{1/2} \tag{5.6}$$

is called a *Littlewood–Paley g-function operator* associated with ψ. It is clear that

$$\begin{aligned}\int |g(f)(\mathbf{x})|^2 d\mathbf{x} &= \frac{1}{(2\pi)^s}\int_0^\infty \frac{dt}{t}\int_{\mathbb{R}^s} |\widehat{\psi}_t(\mathbf{x})\widehat{f}(\mathbf{x})|^2 d\mathbf{x} \\ &= \frac{1}{(2\pi)^s}\int_{\mathbb{R}^s} |\widehat{f}(\mathbf{x})|^2 \left(\sum_{j\in\mathbb{Z}}\int_{2^j}^{2^{j+1}} \frac{dt}{t}|\widehat{\psi}(t\mathbf{x})|^2\right) d\mathbf{x}.\end{aligned}$$

Hence we have

$$\int |g(f)(\mathbf{x})|^2 d\mathbf{x} \le \frac{1}{(2\pi)^s}\int_{\mathbb{R}^s} |\widehat{f}(\mathbf{x})|^2 \int_1^2 \sum_{j\in\mathbb{Z}} |\widehat{\psi}(2^{-j}t\mathbf{x})|^2 dt d\mathbf{x} \le 2\int |g(f)(\mathbf{x})|^2 d\mathbf{x}.$$

and this yield the following result.

Proposition 4. *The g-function operator* (5.6) *is bounded in L^2 if and only if the corresponding Littlewood–Paley energy function*

$$L_\psi(\mathbf{x}) := \sum_{j\in\mathbb{Z}} |\widehat{\psi}(2^j\mathbf{x})|^2$$

satisfies $\int_1^2 L_\psi(t\mathbf{x})dt \in L^\infty$.

By combining this result and Theorem 1, we obtain the following.

Proposition 5. *Let Φ be a nonincreasing function on* $[0,\infty)$ *that satisfies* (1.9) *and ψ satisfies $\int \psi = 0$ and $|\psi(\mathbf{x})| \le \Phi(|\mathbf{x}|)$, $\mathbf{x} \in \mathbb{R}^s$. Then the g-function operator associated with ψ is bounded in L^2.*

REFERENCES

[1] C. K. Chui and X. L. Shi, *Inequalities of Littlewood–Paley type for frames and wavelets*, SIAM J. Math. Anal., 24 (1993), pp. 263–277.

[2] C. K. Chui and X. L. Shi, *Continuous two-scale equations and dyadic wavelets*, Adv. Comput. Math., 2 (1994), pp. 185–213.

[3] C. K. Chui and X. L. Shi, *Bessel sequences and affine frames*, Appl. Comput. Harmonic Anal., 1 (1993), pp. 29–49.

[4] I. Daubechies, *Ten Lectures on Wavelets*, CBMS–NSF Series in Applied Mathematics 61, SIAM, Philadelphia, 1992.

[5] S. Mallat and S. Zhong, *Wavelet transform maxima and multiscale edges*, in Wavelets and Their Applications, M. B. Ruskia, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., Jones and Bartlett, Boston, 1992, pp. 67–104.

[6] Y. Meyer, *Ondelettes et Operateurs*, Vols. I and II, Hermann, Paris, 1990.

[7] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 28, No. 1, pp. 233–258, January 1997

A MULTIVARIATE FORM OF HARDY'S INEQUALITY AND L_p-ERROR BOUNDS FOR MULTIVARIATE LAGRANGE INTERPOLATION SCHEMES*

SHAYNE WALDRON[†]

Abstract. The *multivariate* generalization of *Hardy's inequality*—that for $m - n/p > 0$,

$$(*) \qquad \left\| x \mapsto \int_{[\underbrace{x,\ldots,x}_{m},\Theta]} f \right\|_p \le \frac{\|f\|_p}{(m-1)!(m-n/p)_{\#\Theta}},$$

valid for $f \in L_p(\mathbb{R}^n)$ and Θ an arbitrary finite sequence of points in $\mathbb{R}^n$—is discussed.

The linear functional $f \mapsto \int_\Theta f$ was introduced by Micchelli in connection with Kergin interpolation. This functional also naturally occurs in other multivariate generalizations of Lagrange interpolation, including *Hakopian interpolation* and the *Lagrange maps* of section 5. For each of these schemes, $(*)$ implies L_p-error bounds.

We discuss why $(*)$ plays a crucial role in obtaining L_p-bounds from pointwise integral error formulas for multivariate generalizations of Lagrange interpolation.

Key words. Hardy's inequality, Lagrange interpolation, Kergin interpolation, Hakopian interpolation, B-spline, simplex spline, Hermite–Genocchi formula

AMS subject classifications. Primary, 26D10, 41A10, 41A17, 41A63; Secondary, 41A05, 41A80

PII. S0036141094275506

1. Introduction.

1.1. Overview. The central result of this paper is the following inequality: for $m - n/p > 0$,

$$(1.1.1) \quad \left\| x \mapsto \int_{[\underbrace{x,\ldots,x}_{m},\Theta]} f \right\|_{L_p(\Omega)} \le \frac{1}{(m-1)!(m-n/p)_{\#\Theta}} \|f\|_{L_p(\Omega)}, \quad \forall f \in L_p(\Omega),$$

where the functional $\int_{[x,\cdots,x,\Theta]}$ is defined in Definition 2.1.1, Θ is a finite sequence of points in $\mathbb{R}^n$, and Ω is a suitable domain in $\mathbb{R}^n$. This inequality is a *multivariate* generalization of *Hardy's inequality*—that for $p > 1$,

$$(1.1.2) \qquad \left\| x \mapsto \frac{1}{x}\int_0^x f \right\|_{L_p(0,\infty)} \le \frac{p}{p-1} \|f\|_{L_p(0,\infty)}, \quad \forall f \in L_p(0,\infty).$$

Thus, we will refer to (1.1.1) as the *multivariate form of Hardy's inequality.*

*Received by the editors October 11, 1994; accepted for publication (in revised form) October 10, 1995. This work was supported by the Chebyshev professorship of Carl de Boor and was completed while at the Center for the Mathematical Sciences at the University of Wisconsin–Madison.

http://www.siam.org/journals/sima/28-1/27550.html.

[†]Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand (waldron@math.auckland.ac.nz, http://www.math.auckland.ac.nz/~waldron).

Our interest in (1.1.1) comes from a desire to obtain L_p-bounds from the many integral error formulas for *multivariate* generalizations of Lagrange interpolation that involve the linear functional

$$f \mapsto \int_{[\underbrace{x,\ldots,x}_{m},\Theta]} f. \tag{1.1.3}$$

The paper is structured in the following way. In the remainder of this section, the notation that we will use and the facts about Sobolev spaces that we will need are discussed. In section 2, some properties of the linear functional $f \mapsto \int_\Theta f$, and its connection with simplex splines, are given. In section 3, the multivariate form of Hardy's inequality is proved. In section 4, the multivariate form of Hardy's inequality is applied to obtain L_p-bounds for the error in the scale of mean value interpolations, which includes Kergin and Hakopian interpolation. In section 5, in a similar vein, L_p-bounds for the error in *Lagrange maps* are obtained. In section 6, we discuss why the multivariate form of Hardy's inequality is applicable to the many error formulas for multivariate Lagrange interpolation schemes and is likely to be so for others yet to be obtained.

1.2. Some notation. The discussion takes place in $\mathbb{R}^n$ with the following definitions holding throughout. The space of n-variate polynomials of degree k will be denoted by $\Pi_k(\mathbb{R}^n)$, and the space of homogeneous polynomials of degree k will be denoted by $\Pi_k^0(\mathbb{R}^n)$. The differential operator induced by $q \in \Pi_k(\mathbb{R}^n)$ will be written $q(D)$. Let $\|\cdot\|$ be the *Euclidean norm* on $\mathbb{R}^n$, and let $\Omega \subset \mathbb{R}^n$, with $\bar{\Omega}$ its closure. The letters i, j, k, l, m, n will be reserved for integers, and $1 \le p \le \infty$. We use standard multivariate notation, so, e.g., $\{\alpha : |\alpha| = k\}$ is the set of multiindices α of length k.

We find it convenient to make no distinction between the matrix $[\theta_1, \ldots, \theta_k]$ and the *k-sequence* $\theta_1, \ldots, \theta_k$ of its columns. Since $[\theta_1, \ldots, \theta_k]f$ is a standard notation for the *divided difference* of f at $\Theta = [\theta_1, \ldots, \theta_k]$, we use for the latter the nonstandard notation

$$\delta_\Theta f = \delta_{[\theta_1,\ldots,\theta_k]} f.$$

Note the special case

$$\delta_{[x]} f = f(x).$$

Similarly, to avoid any confusion, the closed interval with endpoints a and b will be denoted by $[a \,..\, b]$.

The derivative of f in the directions Θ is denoted

$$D_\Theta f := D_{\theta_1} \cdots D_{\theta_k} f.$$

The notation $\tilde{\Theta} \subset \Theta$ means that $\tilde{\Theta}$ is a subsequence of Θ, and $\Theta \backslash \tilde{\Theta}$ denotes the complementary subsequence. The subsequence consisting of the first j terms of Θ is denoted Θ_j, and

$$x - \Theta := [x - \theta_1, \ldots, x - \theta_k].$$

Thus, with $\Theta := [\theta_1, \ldots, \theta_7]$, we have, for example, that

$$D_{[x-\Theta\backslash\Theta_5, x-\theta_3]} f = D_{x-\theta_6} D_{x-\theta_7} D_{x-\theta_3} f.$$

The diameter and convex hull of a sequence Θ will be that of the corresponding set and will be denoted by $\operatorname{diam}\Theta$ and $\operatorname{conv}\Theta$, respectively.

Many of the constants in this paper involve the *shifted factorial function*

$$(a)_n := (a)(a+1)(a+2)\cdots(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)}, \tag{1.2.1}$$

where Γ is the *gamma function.* The gamma function satisfies the relation

$$\Gamma(a+1) = a\Gamma(a), \quad \forall a > 0,$$

and has $\Gamma(1) = 1$. In particular,

$$\Gamma(n+1) = n!, \quad n = 0, 1, 2, \ldots. \tag{1.2.2}$$

Some of our calculations require the *beta integrals*

$$\int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \qquad a, b > 0, \tag{1.2.3}$$

and the *hypergeometric function*

$${}_2F_1\left({a, b \atop c}; x\right) := \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{n!(c)_n} x^n. \tag{1.2.4}$$

The standard reference to these is the monograph [E53].

1.3. Geometry of the domain Ω. We say that $\Omega \subset \mathbb{R}^n$ is *star-shaped with respect to S*, a set (resp. sequence) in $\mathbb{R}^n$, when Ω contains the convex hull of $S \cup \{x\}$ for any $x \in \Omega$. This condition is weaker than Ω being convex.

In our results, it will be required that $\bar{\Omega}$ be star-shaped with respect to $\Theta \in \mathbb{R}^{n\times k}$, where Ω is an open set in $\mathbb{R}^n$. This condition is required of $\bar{\Omega}$, rather than of Ω, so as to include cases where some points in Θ lie on the boundary of Ω. (See Figure 1.1.) One such example of interest is the *Lagrange* finite element given by linear interpolation at Θ, the vertices of a n-simplex; see, e.g., Ciarlet [Ci78, p. 46]. In this case, $\bar{\Omega} = \operatorname{conv}\Theta$, and none of the points of Θ are in the open simplex Ω.

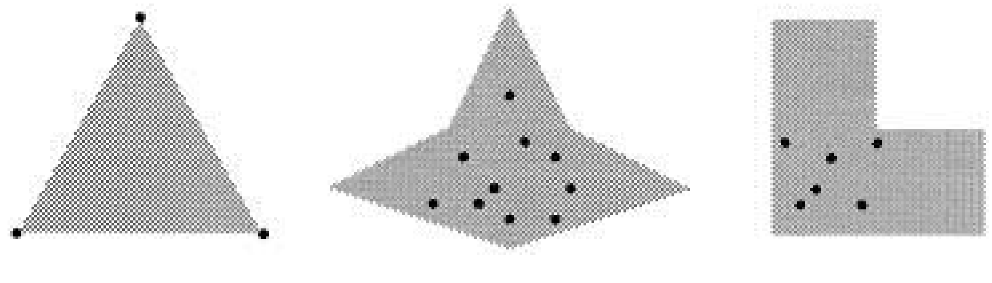


FIG. 1.1. *Examples of domains Ω (shaded) for which $\bar{\Omega}$ is star-shaped with respect to the points in Θ ($\bullet$).*

We now show that being star-shaped with respect to a finite sequence is equivalent to being star-shaped with respect to its convex hull.

PROPOSITION 1.3.1. *If* $\Omega \subset \mathbb{R}^n$ *and* $\Theta \in \mathbb{R}^{n\times k}$*, then the following are equivalent:*

(a) Ω *is star-shaped with respect to* Θ*.*

(b) Ω *is star-shaped with respect to* $\operatorname{conv}\Theta$*.*

Proof. (See Figure 1.2.) Only the implication (a) $\Longrightarrow$ (b) requires proof. Suppose (a). To obtain (b), it suffices to prove that if Ω is star-shaped with respect to points u and v, then $\operatorname{conv}\{u,v,x\} \subset \Omega$, $\forall x \in \Omega$, i.e., Ω is star-shaped with respect to $\operatorname{conv}\{u,v\}$.

Assume without loss of generality that u, v, x are affinely independent, and $z \in \operatorname{conv}\{u,v,x\}$. Let w be the point of intersection of the line through u and z with the interval $\operatorname{conv}\{x,v\}$. Since Ω is star-shaped with respect to v, we have that $w \in \Omega$. Thus, since Ω is star-shaped with respect to u, we have that $z \in \operatorname{conv}\{u,w\} \subset \Omega$. □

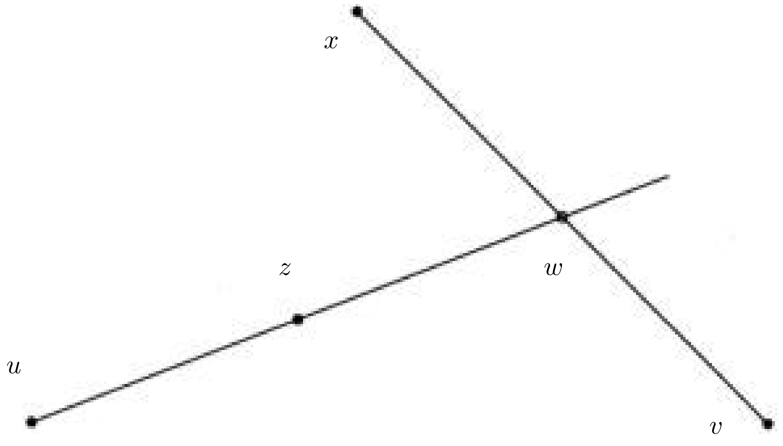


FIG. 1.2. *The proof of Proposition* 1.3.1.

This equivalence ensures that if $\bar{\Omega}$ is star-shaped with respect to Θ, then $f \in L_p(\Omega)$ is defined over the region of integration in (1.1.3) for all $x \in \Omega$.

1.4. Sobolev spaces. Let $W_p^{(k)}(\Omega)$ be the *Sobolev space* consisting of those functions defined on Ω (a bounded open set in $\mathbb{R}^n$ with a *Lipschitz* boundary) with derivatives up to order k in $L_p(\Omega)$, and equipped with the usual topology; see, e.g., Adams [Ad75]. It is convenient to include in the definition the condition that Ω have a Lipschitz boundary so that Sobolev's embedding theorem can be applied. The full statement of Sobolev's embedding theorem can be found in any text on Sobolev spaces, see, e.g., [Ad75, p. 97]; however, we will need only the following consequence of it. If $j - n/p > 0$, then

$$W_p^{(k+j)}(\Omega) \subset C^k(\bar{\Omega}).$$

To measure the size of its kth derivative, it is convenient to associate with each $f \in W_p^{(k)}(\Omega)$ the function $|D^k f| \in L_p(\Omega)$, given by the rule

$$|D^k f|(x) := \sup_{\substack{\Theta\in\mathbb{R}^{n\times k}\\ \|\theta_i\|\le 1}} |D_\Theta f(x)| = \sup_{\substack{\theta\in\mathbb{R}^n\\ \|\theta\|=1}} |D_\theta^k f(x)|, \tag{1.4.1}$$

where the derivatives $D_\Theta f$ are computed from any (fixed) choice of representatives for the partial derivatives $D^\alpha f \in L_p(\Omega)$, $|\alpha| = k$. The equality of the two suprema follows from a classical result of Banach on the norm of a symmetric multilinear map (cf. Harris [Har96]). It is also proved in Chen and Ditzian [CD90]. This definition of $|D^k f|$ is consistent with its standard univariate interpretation. From (1.4.1), it is easy to see that $|D^k f|$ is well defined and satisfies

$$|D_\Theta f| \le |D^k f| \, \|\theta_1\| \cdots \|\theta_k\| \tag{1.4.2}$$

for all $\Theta \in \mathrm{I\!R}^{n\times k}$. The inequality (1.4.2) holds a.e. To emphasize that $D_\Theta f$ and $|D^k f|$ are in $L_p(\Omega)$, we will say that (1.4.2) holds in $L_p(\Omega)$. The $L_p(\Omega)$-norm of $|D^k f|$ gives a seminorm for $f \in W_p^{(k)}(\Omega)$,

$$f \mapsto \big|f\big|_{k,p,\Omega} := \| \, |D^k f| \, \|_{L_p(\Omega)}. \tag{1.4.3}$$

Because of (1.4.2), this coordinate-independent seminorm (1.4.3) is more appropriate for the analysis that follows than other equivalent seminorms, such as

$$f \mapsto \| \, (\|D^\alpha f\|_{L_p(\Omega)} : |\alpha| = k) \, \|_p.$$

2. The linear functional $f \mapsto \int_\Theta f$.

2.1. Definitions. The construction of the maps of Kergin and Hakopian depends intimately on the following linear functional, called the *divided difference functional on* $\mathrm{I\!R}^n$ by Micchelli in [M79], and analyzed there and in [M80].

Definition 2.1.1. *For any $\Theta \in \mathrm{I\!R}^{n\times(k+1)}$, let*

$$f \mapsto \int_\Theta f := \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-1}} f(\theta_0 + s_1(\theta_1 - \theta_0) + \cdots + s_k(\theta_k - \theta_{k-1})) \, ds_k \cdots ds_2 \, ds_1$$

with the convention that $\int_{[]} f := 0$.

In addition to Kergin and Hakopian interpolation, the linear functional $f \mapsto \int_\Theta f$ also occurs when discussing other *multivariate* generalizations of Lagrange interpolation, e.g., the *Lagrange maps* of section 5. It was used as early as 1878, when in [Ge1878], Genocchi proved the *Hermite–Genocchi formula*, namely, that for $\Theta \in \mathrm{I\!R}^{1\times(k+1)}$ and $f \in C^k(\mathrm{conv}\,\Theta)$,

$$\delta_\Theta f = \int_\Theta D^k f.$$

In this section, we outline those properties of $f \mapsto \int_\Theta f$ needed in the subsequent sections. Many of these properties are apparent from the following observation.

Observation 2.1.2. *If S is any k-simplex in $\mathrm{I\!R}^m$ and $A : \mathrm{I\!R}^m \to \mathrm{I\!R}^n$ is any affine map taking the $k+1$ vertices of S onto the $k+1$ points in Θ, then*

$$\int_\Theta f = \frac{1}{k! \, \mathrm{vol}_k(S)} \int_S f \circ A,$$

with $\mathrm{vol}_k(S)$ *the* (k*-dimensional*) *volume of* S.

With the choice

$$A : \mathbb{R}^k \to \mathbb{R}^n : (s_1, \dots, s_k) \mapsto \theta_0 + s_1(\theta_1 - \theta_0) + \cdots + s_k(\theta_k - \theta_{k-1}),$$

$$S := \{(s_1, \dots, s_k) \in \mathbb{R}^k : 0 \le s_k \le \cdots \le s_2 \le s_1 \le 1\},$$

this is simply Definition 2.1.1. The different choice

$$A : \mathbb{R}^{k+1} \to \mathbb{R}^n : (v_0, \dots, v_k) \mapsto v_0\theta_0 + \cdots + v_k\theta_k,$$

$$S := \left\{(v_0, \dots, v_k) \in \mathbb{R}^{k+1} : v_j \ge 0,\ \sum_{j=0}^{k} v_j = 1\right\}$$

shows that our definition of $\int_\Theta f$ coincides with the one used by Micchelli in [M80].

Properties 2.1.3.

(a) *The value of* $\int_\Theta f$ *does not depend on the ordering of the points in* Θ*.*

(b) *The distribution*

$$M_\Theta : C_0^\infty(\mathbb{R}^n) \to \mathbb{R} : f \mapsto k! \int_\Theta f$$

is the (normalized) simplex spline with knots Θ (*cf.* [M80])*.*

(c) *If* $f \in C(\operatorname{conv}\Theta)$*, then* $\int_\Theta f$ *is defined, and for some* $\xi \in \operatorname{conv}\Theta$*,*

$$\int_\Theta f = \frac{1}{k!} f(\xi).$$

(d) *If* $g : \mathbb{R}^s \to \mathbb{R}$ *and if* $B : \mathbb{R}^n \to \mathbb{R}^s$ *is an affine map, then*

$$\int_\Theta (g \circ B) = \int_{B\Theta} g.$$

Remark 2.1.4. Let $A|$ denote the restriction of A to the orthogonal complement of its kernel, which is a 1–1 map onto the affine hull of Θ. The simplex spline M_Θ of (b) has support $\operatorname{conv}\Theta$. It can be represented by the nonnegative bounded function

$$\operatorname{conv}\Theta \to \mathbb{R} : t \mapsto M(t|\Theta) := \frac{\operatorname{vol}_{k-d}(A^{-1}t \cap S)}{|\det(A|)| \operatorname{vol}_k(S)}, \quad d := \dim \operatorname{conv}\Theta,$$

in the sense that

$$M_\Theta f = \int_{\operatorname{conv}\Theta} M(\cdot|\Theta) f. \tag{2.1.5}$$

In particular, if the points of Θ are affinely independent, then

$$k! \int_\Theta f = \frac{1}{\operatorname{vol}_k(\operatorname{conv}\Theta)} \int_{\operatorname{conv}\Theta} f = \text{average value of } f \text{ on } \operatorname{conv}\Theta. \tag{2.1.6}$$

Thus $\int_\Theta f$ is defined (as a real number) if and only if $M(\cdot|\Theta) f \in L_1(\operatorname{conv}\Theta)$, in which case

$$\left|\int_\Theta f\right| \le \int_\Theta |f|. \tag{2.1.7}$$

If f is nonnegative on $\operatorname{conv}\Theta$, then $\int_\Theta f \in [0\mathinner{..}\infty]$ is defined (by Definition 2.1.1). Therefore, we will write (2.1.7) for all f which are defined on $\operatorname{conv}\Theta$—with the understanding that $\int_\Theta f$ is defined if and only if $\int_\Theta |f| < \infty$ or f is nonnegative. In the univariate case, that is, when $n = 1$, $M(\cdot|\Theta)$ is the (*normalized*) *B-spline* with knots Θ. For additional details about M_Θ and $M(\cdot|\Theta)$, see, e.g., Micchelli [M79].

Example 2.1.8. As a special case of (2.1.5), we have

$$\int_{[\underbrace{0,\dots,0}_{m},\ \underbrace{1,\dots,1}_{k+1-m}]} f = \frac{1}{(m-1)!(k-m)!}\int_0^1 t^{k-m}(1-t)^{m-1} f(t)\,dt.$$

Thus, by Property 2.1.3(d) with $B : t \mapsto x + t(v-x)$ and $\Theta = [0,\dots,0,1,\dots,1]$,

$$\begin{aligned}\int_{[\underbrace{x,\dots,x}_{m},\ \underbrace{v,\dots,v}_{k+1-m}]} f &= \int_{[\underbrace{0,\dots,0}_{m},\ \underbrace{1,\dots,1}_{k+1-m}]} f(x+\cdot(v-x)) \\ &= \frac{1}{(m-1)!(k-m)!}\int_0^1 t^{k-m}(1-t)^{m-1} f(x+t(v-x))\,dt.\end{aligned} \tag{2.1.9}$$

2.2. Some technical details.

Remark 2.2.1. In view of Properties 2.1.3(a),

$$\Theta \mapsto \int_\Theta f$$

could be thought of as a map defined on finite multisets in $\mathbb{R}^n$ rather than on sequences. However, adopting this definition leads to certain unnecessary complications. For example, to discuss the continuity of $\Theta \mapsto \int_\Theta f$, it would be necessary to endow the set of multisets of $k+1$ points in $\mathbb{R}^n$ with the appropriate topology. Thus, in the interest of simplicity, $\Theta \mapsto \int_\Theta f$ remains a map on sequences—but with the reader encouraged to think of it, as does the author, as a map on multisets.

Finally, by (2.1.5), we can describe the continuity of $\Theta \mapsto \int_\Theta f$ as follows.

PROPOSITION 2.2.2.

(a) *For $f \in C(\mathbb{R}^n)$, the map*

$$\mathbb{R}^{n\times(k+1)} \to \mathbb{R} : \Theta \mapsto \int_\Theta f$$

is continuous.

(b) *For $f \in L_1^{\mathrm{loc}}(\mathbb{R}^n)$, the map*

$$\{\Theta \in \mathbb{R}^{n\times(k+1)} : \operatorname{vol}_n(\operatorname{conv}\Theta) > 0\} \to \mathbb{R} : \Theta \mapsto \int_\Theta f$$

is continuous.

3. The main results: The multivariate form of Hardy's inequality and L_p-inequalities.

In this section, we prove the multivariate form of Hardy's inequality. This inequality is useful for obtaining L_p-bounds from integral error formulas for various multivariate interpolation schemes.

First, we need a technical lemma.

3.1. A lemma.

LEMMA 3.1.1. *Let m, k, be integers, and $\mu \in \mathbb{R}$. If $1 \le m \le k$ and $m+\mu > 0$, then*

$$\int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-1}} (1-s_m)^\mu \, ds_k \cdots ds_1 = \frac{1}{(m-1)!(m+\mu)_{k+1-m}}.$$

Proof. This can be proved by successively evaluating the univariate integrals. Instead, a proof using the properties of $f \mapsto \int_\Theta f$ is given.

From Definition 2.1.1, it is seen that

$$\int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-1}} (1-s_m)^\mu \, ds_k \cdots ds_1 = \int_\Theta (\cdot)^\mu,$$

where

$$\Theta := [\underbrace{1,\ldots,1}_{m}, \underbrace{0,\ldots,0}_{k+1-m}].$$

By (2.1.9), (1.2.3), and (1.2.2), it follows that

$$\begin{aligned} \int_\Theta (\cdot)^\mu &= \frac{1}{(m-1)!(k-m)!} \int_0^1 t^{k-m}(1-t)^{m-1}(1-t)^\mu \, dt \\ &= \frac{1}{(m-1)!(k-m)!} \frac{\Gamma(k-m+1)\Gamma(m+\mu)}{\Gamma(k+1+\mu)} \\ &= \frac{1}{(m-1)!(m+\mu)_{k+1-m}}. \end{aligned}$$

Here the condition that $m+\mu > 0$ is needed to ensure that the beta integral is finite. □

3.2. The multivariate form of Hardy's inequality. We now prove the multivariate form of Hardy's inequality.

THEOREM 3.2.1. *Let Θ be a nonempty finite sequence in $\mathbb{R}^n$, and let Ω be an open set in $\mathbb{R}^n$ for which $\bar{\Omega}$ is star-shaped with respect to Θ. If $m - n/p > 0$, then the rule*

$$H_{m,\Theta} f(x) := \int_{[\underbrace{x,\ldots,x}_{m},\Theta]} f \tag{3.2.2}$$

induces a positive bounded linear map $H_{m,\Theta} : L_p(\Omega) \to L_p(\Omega)$ with norm

$$\|H_{m,\Theta}\|_{L_p(\Omega)} \le \frac{1}{(m-1)!(m-n/p)_{\#\Theta}} \to \infty \qquad \text{as } m - n/p \to 0^+. \tag{3.2.3}$$

This upper bound for $\|H_{m,\Theta}\|_{L_p(\Omega)}$ is sharp when Θ involves only one point, i.e., when

$$\Theta = [v, \ldots, v],$$

and is also sharp when $p = \infty$. Furthermore, if $\Omega \subset \Omega'$, then

$$\|H_{m,\Theta}\|_{L_p(\Omega)} \le \|H_{m,\Theta}\|_{L_p(\Omega')}. \tag{3.2.4}$$

Proof. Suppose that $m - n/p > 0$. Then $m > 0$. Let $k+1 := m + \#\Theta$, and write

$$[\underbrace{x,\ldots,x}_{m},\Theta] =: [\underbrace{x,\ldots,x}_{m},\theta_m,\theta_{m+1},\ldots,\theta_k].$$

By Definition 2.1.1,

$$H_{m,\Theta}f(x) = \int_S f(A_x s)\,ds, \tag{3.2.5}$$

where $s := (s_1,\ldots,s_k)$,

$$\int_S := \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-1}}, \qquad ds := ds_k \cdots ds_1,$$

and

$$A_x s := x + s_m(\theta_m - x) + s_{m+1}(\theta_{m+1} - \theta_m) + \cdots + s_k(\theta_k - \theta_{k-1}).$$

The domain of integration for f in (3.2.5) is $\operatorname{conv}[x,\Theta]$, which by Proposition 1.3.1 is contained (up to a set of measure zero) in Ω for any $x \in \Omega$. However, for $f \in L_p(\Omega)$, it is not clear whether the integrals in (3.2.5) converge so as to define a function $H_{m,\Theta}f$ which is in $L_p(\Omega)$ (or is even measurable for that matter).

First, suppose that f is a *nonnegative measurable function.* Then (3.2.5) defines a nonnegative measurable function $H_{m,\Theta}f$, as is now shown. The nonnegativity of $H_{m,\Theta}f$, i.e., the positiveness of the map $H_{m,\Theta}$, is obvious, and the measurability of $H_{m,\Theta}f$ is a consequence of Tonelli's theorem (see, e.g., Folland [Fo84, p. 65]) as follows.

First, we prove that the map

$$(x,s) \mapsto f(A_x s) \tag{3.2.6}$$

is measurable. Since f is measurable, the measurability of (3.2.6) is equivalent to $A^{-1}(E)$ being measurable for each $E \in \mathcal{E}$, where

$$A : (x,s) \mapsto A_x s,$$

and $\mathcal{E}$ is any family of sets that generates the Lebesgue σ-algebra. Take $\mathcal{E}$ as the Borel sets together with all the subsets F of any Borel set B of measure zero. Since A is continuous, the inverse image under A of a Borel set is a Borel set (which is measurable). For $F \subset B$, $A^{-1}(F)$ is contained within the Borel set $A^{-1}(B)$, which has zero measure (see below), and hence is measurable. For $s_m \neq 1$, the set

$$\{x : A_x s \in B\} = \frac{1}{1-s_m}(B - s_m\theta_m - s_{m+1}(\theta_{m+1}-\theta_m) - \cdots - s_k(\theta_k - \theta_{k-1})),$$

hence has zero measure, and so, by Tonelli's theorem,

$$\operatorname{meas}(A^{-1}(B)) = \int_S \operatorname{meas}(\{x : A_x s \in B\})\,ds = 0.$$

This completes the proof of the measurability of (3.2.6). Since (3.2.6) is a nonnegative measurable function, it follows from Tonelli's theorem that

$$H_{m,\Theta} : x \mapsto \int_S f(A_x s)$$

is measurable.

Apply Minkowski's inequality for integrals (see, e.g., Folland [Fo84, p. 186]) to the sum (integral) $\int_S$ of functions $x \mapsto f(A_x s)$ to obtain

$$\|H_{m,\Theta} f\|_{L_p(\Omega)} \le \int_S \|x \mapsto f(A_x s)\|_{L_p(\Omega)}\, ds. \tag{3.2.7}$$

The case where $1 \le p < \infty$. Inequality (3.2.7) can be written as

$$\|H_{m,\Theta} f\|_{L_p(\Omega)} \le \int_S \left(\int_\Omega f(A_x s)^p \, dx \right)^{1/p} ds.$$

Make the change of variables

$$y = A_x s$$

in the inner integral above. The new region of integration is contained in Ω, and $dy = (1 - s_m)^n dx$. Thus, by the change of variables formula (see, e.g., Rudin [Ru87, p. 153]), it follows that

$$\begin{aligned} \int_S \left(\int_\Omega f(A_x s)^p \, dx \right)^{1/p} ds &\le \int_S \left(\int_\Omega \frac{f(y)^p \, dy}{(1 - s_m)^n} \right)^{1/p} ds \\ &= \left(\int_S (1 - s_m)^{-n/p} \, ds \right) \|f\|_{L_p(\Omega)}. \end{aligned}$$

From Lemma 3.1.1 with $k + 1 - m = \#\Theta$ and $\mu = -n/p$, it follows that

$$\int_S (1 - s_m)^{-n/p} \, ds = \frac{1}{(m-1)!(m - n/p)_{\#\Theta}}. \tag{3.2.8}$$

The case where $p = \infty$. Since $x \mapsto A_x s$ maps sets of measure zero to sets of measure zero, it follows from (3.2.7) that

$$\|H_{m,\Theta} f\|_{L_\infty(\Omega)} \le \int_S \|f\|_{L_\infty(\Omega)} \, ds = \frac{1}{k!} \|f\|_{L_\infty(\Omega)} \tag{3.2.9}$$

with equality when f is constant. The fact that

$$\int_S ds = \frac{1}{k!},$$

used above, follows from Observation 2.1.2 or from Lemma 3.1.1 with $\mu = 0$.

So far, it has been shown that for a nonnegative measurable f, (3.2.2) defines a nonnegative measurable function which satisfies

$$\|H_{m,\Theta} f\|_{L_p(\Omega)} \le \frac{1}{(m-1)!(m - n/p)_{\#\Theta}} \|f\|_{L_p(\Omega)}. \tag{3.2.10}$$

In view of this inequality, $H_{m,\Theta}$ induces a map from the nonnegative functions in $L_p(\Omega)$ to $L_p(\Omega)$. Each $f \in L_p(\Omega)$ can be written as

$$f = f^+ - f^-,$$

a difference of nonnegative functions in $L_p(\Omega)$ (its *positive* and *negative parts*), and so (due to its linearity) $H_{m,\Theta}$ induces a map on $L_p(\Omega)$, also denoted by $H_{m,\Theta}$. Since

$$\|H_{m,\Theta}f\|_{L_p(\Omega)} \le \|H_{m,\Theta}(|f|)\|_{L_p(\Omega)}, \quad \forall f \in L_p(\Omega),$$

inequality (3.2.10) holds for all $f \in L_p(\Omega)$, which gives (3.2.3).

Next, (3.2.4) is shown. Since the restriction map

$$L_p(\Omega') \to L_p(\Omega) : f \mapsto f|_\Omega$$

is onto and $(H_{m,\Theta}f)|_\Omega$ depends only on $f|_\Omega$,

$$\begin{aligned}\|H_{m,\Theta}\|_{L_p(\Omega)} &= \sup_{f\in L_p(\Omega')} \frac{\|H_{m,\Theta}(f|_\Omega)\|_{L_p(\Omega)}}{\|f|_\Omega\|_{L_p(\Omega)}} \le \sup_{\substack{f\in L_p(\Omega')\\ f=\chi_\Omega f}} \frac{\|H_{m,\Theta}f\|_{L_p(\Omega')}}{\|f\|_{L_p(\Omega')}} \\ &\le \sup_{f\in L_p(\Omega')} \frac{\|H_{m,\Theta}f\|_{L_p(\Omega')}}{\|f\|_{L_p(\Omega')}} = \|H_{m,\Theta}\|_{L_p(\Omega')}.\end{aligned}$$

Finally, the sharpness is proved. Suppose that $\Theta = [v, \ldots, v]$. Let

$$f := \|\cdot - v\|^\alpha, \quad \alpha \in \mathbb{R}.$$

Then by (2.1.9), and (1.2.3), for $m + \alpha > 0$,

$$\begin{aligned}H_{m,\Theta}f(x) &= \frac{1}{(m-1)!(\#\Theta-1)!}\int_0^1 t^{\#\Theta-1}(1-t)^{m-1}\,\|x + t(v-x) - v\|^\alpha\,dt \\ &= \frac{1}{(m-1)!(\#\Theta-1)!}\int_0^1 t^{\#\Theta-1}(1-t)^{m-1+\alpha}\,dt\,\|x-v\|^\alpha \\ &= \frac{1}{(m-1)!(\#\Theta-1)!}\frac{\Gamma(\#\Theta)\Gamma(m+\alpha)}{\Gamma(\#\Theta+m+\alpha)}\,\|x-v\|^\alpha \\ &= \frac{1}{(m-1)!(m+\alpha)_{\#\Theta}}\,\|x-v\|^\alpha,\end{aligned}$$

so that $f := \|\cdot - v\|^\alpha$, $m + \alpha > 0$, is an eigenvector of $H_{m,\Theta}$ with eigenvalue

$$\lambda := \frac{1}{(m-1)!(m+\alpha)_{\#\Theta}}.$$

Thus

$$\begin{aligned}\|H_{m,\Theta}\|_{L_p(\Omega)} &\ge \sup\left\{\frac{1}{(m-1)!(m+\alpha)_{\#\Theta}} : \|\cdot - v\|^\alpha \in L_p(\Omega),\ \alpha + m > 0\right\} \\ &\ge \sup\left\{\frac{1}{(m-1)!(m+\alpha)_{\#\Theta}} : \alpha > -n/p\right\} \\ &= \frac{1}{(m-1)!(m-n/p)_{\#\Theta}},\end{aligned}$$

giving equality in (3.2.3). The sharpness for the case where $p = \infty$ follows from the observation that there is sharpness in inequality (3.2.9) for f constant and Ω bounded, together with inequality (3.2.4). □

Remark 3.2.11. If $\mathrm{vol}_n(\mathrm{conv}\,\Theta) > 0$, then by Remark 2.1.4, it follows that the value of $H_{m,\Theta}f(x)$ is the same for all representatives of $f \in L_p(\Omega)$. Indeed, by Proposition 2.2.2, for all $f \in L_p(\Omega)$, we have that $H_{m,\Theta}f \in C(\bar{\Omega})$, regardless of whether or not $m - n/p > 0$.

On the other hand, when $\mathrm{vol}_n(\mathrm{conv}\,\Theta) = 0$, then the function $H_{m,\Theta}f$ need not be so well behaved. For example, if $n > 1$ and Θ consists of a single point θ, then $f \in L_p(\Omega)$ can be altered on a null set so that $H_{m,\Theta}f$ takes on arbitrary preassigned values on any countable dense subset of Ω. For the details of one such construction, see the end of this section.

3.3. Special case: Hardy's inequality. In the very special case where $n = 1$, $m = 1$, and $\Theta = [0]$, we have by (2.1.6) that

$$H_{m,\Theta}f(x) = \frac{1}{x}\int_0^x f. \tag{3.3.1}$$

With the choice $\Omega = (0, \infty)$, (3.2.3) is Hardy's inequality (1.1.2). This well-known inequality was first proved by Hardy [Ha28]; see also [HLP67, section 9.8].

For a comprehensive survey of the literature connected with Hardy's inequality, see [FMP91, Chapter IV]. The only *multivariate* occurrence of Theorem 3.2.1 that the author is aware of is, implicitly, in Arcangeli and Gout [AG76] for the case when Θ consists of a single point. The bulk of the 174 references for [FMP91, Chapter IV] deal with *univariate* generalizations of Hardy's inequality—some of which are special cases of Theorem 3.2.1.

3.4. Further L_p-bounds. Next, we use Theorem 3.2.1 to give a bound particularly suited for obtaining L_p-bounds from integral error formulas, such as those given in sections 4 and 5.

THEOREM 3.4.1. *Fix $a_1, \dots, a_s \in \mathbb{R}^{k+1} \setminus 0$, where $s \geq 0$. Let $\Theta \in \mathbb{R}^{n \times k}$, and let Ω be a bounded open set in $\mathbb{R}^n$ for which $\bar{\Omega}$ is star-shaped with respect to Θ. If $m - n/p > 0$, then the rule*

$$\mathcal{L}f(x) := \int_{[\underbrace{x,\dots,x}_{m},\Theta]} \left(\prod_{j=1}^{s} D_{[x,\Theta]a_j}\right) f \tag{3.4.2}$$

induces a bounded linear map $\mathcal{L} : W_p^s(\Omega) \to L_p(\Omega)$ with

$$\|\mathcal{L}f\|_{L_p(\Omega)} \leq \left(\max_{x \in \bar{\Omega}} \prod_{j=1}^{s} \|[x,\Theta]a_j\|\right) \frac{1}{(m-1)!(m-n/p)_{\#\Theta}} \left|f\right|_{s,p,\Omega}. \tag{3.4.3}$$

In addition, when $p = \infty$, we have the pointwise estimate

$$|\mathcal{L}f(x)| \leq \frac{1}{(\#\Theta + m - 1)!}\left(\prod_{j=1}^{s} \|[x,\Theta]a_j\|\right)\left|f\right|_{s,\infty,\Omega}, \qquad a.e.\ x \in \Omega. \tag{3.4.4}$$

Proof. It follows from Theorem 3.2.1 that (3.4.2) induces a linear map $W_p^s(\Omega) \to L_p(\Omega)$. Next, (3.4.3) is proved.

Let $x \in \Omega$ and $f \in W_p^s(\Omega)$. By (1.4.2),

$$\left|\left(\prod_{j=1}^{s} D_{[x,\Theta]a_j}\right) f\right| \leq \left(\prod_{j=1}^{s} \|[x,\Theta]a_j\|\right) |D^s f| \tag{3.4.5}$$

in $L_p(\Omega)$. Here $|D^s f| \in L_p(\Omega)$ is defined by (1.4.1). Thus

$$A_x f := \left(\prod_{j=1}^{s} D_{[x,\Theta]a_j}\right) f$$

defines a bounded linear map $A_x : W_p^s(\Omega) \to L_p(\Omega)$ with

$$|A_x f| \leq K\,|D^s f| \tag{3.4.6}$$

in $L_p(\Omega)$, where

$$K := K(a_1, \ldots, a_s, \Omega) := \max_{x \in \bar{\Omega}} \prod_{j=1}^{s} \|[x,\Theta]a_j\|.$$

Notice that

$$\mathcal{L}f(x) = (H_{m,\Theta}\, A_x f)\,(x).$$

Thus, (3.4.6) and the *positiveness* of $H_{m,\Theta} : L_p(\Omega) \to L_p(\Omega)$ implies that

$$|\mathcal{L}f| \leq H_{m,\Theta}(K\,|D^s f|),$$

in $L_p(\Omega)$. Take the $L_p(\Omega)$-norm of this inequality and then apply Theorem 3.2.1 to obtain

$$\|\mathcal{L}f\|_{L_p(\Omega)} \leq \frac{1}{(m-1)!(m-n/p)_{\#\Theta}} K\,\|\,|D^s f|\,\|_{L_p(\Omega)}.$$

Since

$$\|\,|D^s f|\,\|_{L_p(\Omega)} = \bigl|f\bigr|_{s,p,\Omega},$$

this proves (3.4.3).

Similarly, from (3.4.5) and Theorem 3.2.1, we have for a.e. $x \in \Omega$ that

$$\begin{aligned} |\mathcal{L}f(x)| &\leq \left(\prod_{j=1}^{s} \|[x,\Theta]a_j\|\right) \|H_{m,\Theta}(|D^s f|)\|_{L_\infty(\Omega)} \\ &\leq \left(\prod_{j=1}^{s} \|[x,\Theta]a_j\|\right) \frac{1}{(\#\Theta + m - 1)!} \bigl|f\bigr|_{s,\infty,\Omega}, \end{aligned}$$

which is (3.4.4). □

In the special case when $s = 0$, Theorem 3.4.1 reduces to Theorem 3.2.1. Theorem 3.4.1 together with Property 2.1.3(d) can be used to obtain bounds for maps more general than (3.4.2). One such example is the *lift* of an *elementary liftable map*; see [Wa97].

3.5. An example. Finally, here is the example promised in Remark 3.2.11.

Let $n > 1$ and Θ consist of the single point θ. Suppose that $\bar{\Omega}$ is star-shaped with respect to θ, and that B is a countable dense subset of Ω. It is possible to change $f \in L_p(\Omega)$ on the intersection of Ω with the cone C with vertex θ and base B, which is a null set, so that $H_{m,[\theta]}f$, as computed from (3.2.2), takes on arbitrary preassigned values on B.

The cone C consists of the union of rays r emanating from θ and passing through a point $b \in B$. Let r be such a ray, and order the points from B lying on r as $b_1, b_2, \ldots$, so that b_i is closer to θ than b_{i+1}. By Remark 2.1.4,

$$H_{m,[\theta]} f(b_i) = \int M(\cdot | \underbrace{b_i, \ldots, b_i}_{m}, \theta)\, f$$

with the integration above being over the interval $[\theta \mathinner{\ldotp\ldotp} b_i] := \operatorname{conv}\{\theta, b_i\}$ weighted by a nonnegative polynomial. Thus, by redefining f to be an appropriate constant over each of the intervals $[\theta \mathinner{\ldotp\ldotp} b_1]$, $[b_1 \mathinner{\ldotp\ldotp} b_2]$, $[b_2 \mathinner{\ldotp\ldotp} b_3], \ldots$, one can make $H_{m,[\theta]} f(b_i)$ take on any preassigned values.

The function $H_{m,[\theta]} f$ is more than simply an interesting example. It occurs in the *multipoint Taylor error formula* for multivariate Lagrange interpolation given by Ciarlet and Raviart [CR72]. From the multipoint Taylor formula, Arcangeli and Gout [AG76] obtained L_p-bounds for multivariate Lagrange interpolation, long used by those working in finite elements, but known to few approximation theorists. For this reason, these bounds are discussed in some detail in Section 5.

4. Application: L_p-error bounds for Kergin and Hakopian interpolation. In this section, we use Theorem 3.4.1 to obtain L_p-error bounds for the *scale of mean value interpolations*, which includes the *Kergin* and *Hakopian maps*.

To describe the mean-value interpolations and the Lagrange maps of section 5, we will need the following facts about linear interpolation.

4.1. Linear interpolation. Let F be a finite-dimensional space and Λ a finite-dimensional space of linear functionals defined at least on F. We say that the corresponding *linear interpolation problem*, $\operatorname{LIP}(F, \Lambda)$ for short, is *correct* if for every g upon which Λ is defined, there is a unique $f \in F$ which agrees with g on Λ, i.e.,

$$\lambda(f) = \lambda(g), \quad \forall \lambda \in \Lambda.$$

The linear map $L : g \mapsto f$ is called the associated (*linear*) *projector* with *interpolants* F and *interpolation conditions* Λ. Each linear projector with finite-dimensional range F is the solution of a $\operatorname{LIP}(F, \Lambda)$ for some unique choice of the interpolation conditions Λ.

Notice that the correctness of $\operatorname{LIP}(F, \Lambda)$ depends only on the action of Λ on F.

4.2. The scale of mean value interpolations. Throughout this section, $\Theta \in \mathrm{I\!R}^{n \times k}$. For $0 \le m < k$, we have the *mean value interpolation*

$$\mathcal{H}_\Theta^{(m)} : \{f : f \text{ is } C^{k-m-1} \text{ on } \operatorname{conv} \Theta\} \to \Pi_{k-m-1}(\mathrm{I\!R}^n),$$

which is given by

$$\mathcal{H}_\Theta^{(m)} f(x) := m! \sum_{j=m+1}^{k} \sum_{\substack{\tilde\Theta \subset \Theta_{j-1} \\ \#\tilde\Theta = m}} \int_{\Theta_j} D_{x - \Theta_{j-1} \setminus \tilde\Theta} f.$$

$\mathcal{H}_\Theta^{(m)}$ is a linear projector with interpolants $\Pi_{k-m-1}(\mathrm{I\!R}^n)$ and interpolation conditions

$$\operatorname{span}\{f \mapsto \int_{\tilde\Theta} q(D) f : \tilde\Theta \subset \Theta,\ \#\tilde\Theta \ge m+1,\ q \in \Pi^0_{\#\tilde\Theta - m - 1}(\mathrm{I\!R}^n)\}.$$

The map $\mathcal{H}_\Theta^{(0)}$ is *Kergin's map*, and $\mathcal{H}_\Theta^{(n-1)}$ is *Hakopian's map*. The Kergin interpolant matches function values at Θ, as does the Hakopian interpolant in case Θ is in general position; however, this latter fact is not obvious. For this reason, the scale $(\mathcal{H}_\Theta^{(m)} : 0 \le m < k)$ of multivariate mean-value interpolations is thought of as a multivariate generalization of Lagrange interpolation. For more details, see [Wa97].

For the remainder of this section, Ω will be a bounded open set in $\mathbb{R}^n$ with a Lipschitz boundary. From [Wa97], we obtain the following integral error formulas for the scale of mean-value interpolations.

THEOREM 4.2.1. *Suppose that $\bar{\Omega}$ is star-shaped with respect to Θ. If $0 \le j < k-m$, $q \in \Pi_j^0(\mathbb{R}^n)$, $p > n$, and $f \in W_p^{(k-m)}(\Omega)$, then*
(4.2.2)

$$q(D)\bigl(f-\mathcal{H}_\Theta^{(m)} f\bigr)(x) = (m+j)! \sum_{i=k-m-j}^{k} \sum_{\substack{\tilde{\Theta}\subset\Theta_{i-1} \\ \#\tilde{\Theta}=m+j+i-k}} \int_{[\underbrace{x,\dots,x}_{k+1-i},\Theta_i]} D_{[x-\Theta_{i-1}\setminus\tilde{\Theta},x-\theta_i]} q(D) f.$$

This formula involves only derivatives of f of order $k-m$.

Remark 4.2.3. In [Wa97], formula (4.2.2) was proved only for $f \in C^{k-m}(\mathbb{R}^n)$, without any reference to p. We now outline how it can be extended to $f \in W_p^{(k-m)}(\Omega)$. By Sobolev's embedding theorem, the condition $p > n$ implies that

$$W_p^{(k-m)}(\Omega) \subset C^{k-m-1}(\bar{\Omega}) \subset C(\bar{\Omega}).$$

Thus $\mathcal{H}_\Theta^{(m)} f$ is defined for all $f \in W_p^{(k-m)}(\Omega)$. To extend (4.2.2) to $f \in W_p^{(k-m)}(\Omega)$, use a density argument.

4.3. L_p-bounds for the scale of mean-value interpolations. Next, we apply Theorem 3.4.1 to (4.2.2) to obtain L_p-bounds for the scale of mean-value interpolations. Let

$$h_{x,\Theta} := \sup_{\theta\in\Theta} \|x-\theta\|, \qquad h_{\Omega,\Theta} := \sup_{x\in\Omega} h_{x,\Theta} \le \operatorname{diam}\Omega.$$

THEOREM 4.3.1. *Suppose that $\bar{\Omega}$ is star-shaped with respect to Θ. If $0 \le j < k-m$, $p > n$, and $f \in W_p^{(k-m)}(\Omega)$, then*

$$\bigl|f-\mathcal{H}_\Theta^{(m)} f\bigr|_{j,p,\Omega} \le C_{n,p,j,k,m}\,(h_{\Omega,\Theta})^{k-m-j}\bigl|f\bigr|_{k-m,p,\Omega}, \tag{4.3.2}$$

where

$$C_{n,p,j,k,m} := \frac{1}{(1-n/p)_{k-m-j}}.$$

The constant $C_{n,p,j,k,m} \to \infty$ as $p \to n^+$. Additionally, if $p = \infty$, then we have the pointwise estimate that for all $x \in \bar{\Omega}$,

$$|D^j(f-\mathcal{H}_\Theta^{(m)} f)|(x) \le \frac{1}{(k-m-j)!}(h_{x,\Theta})^{k-m-j}\bigl|f\bigr|_{k-m,\infty,\Omega}.$$

Proof. Choose $q \in \Pi_j^0(\mathbb{R}^n)$ so that

$$q(D) = D_{u_1}\cdots D_{u_j},$$

where $u_1, \ldots, u_j \in \mathbb{R}^n$ with $\|u_i\| \le 1$. By Theorem 3.4.1, we have for each of the terms in (4.2.2) that

$$\Bigl\| x \mapsto \int_{[\underbrace{x,\ldots,x}_{k+1-i},\Theta_i]} D_{[x-\Theta_{i-1}\setminus\tilde{\Theta},x-\theta_i]} q(D) f \Bigr\|_{L_p(\Omega)}$$
$$\le \frac{\Gamma(k+1-i-n/p)}{\Gamma(k+1-i)\Gamma(k+1-n/p)} (h_{\Omega,\Theta})^{k-m-j} \bigl| f \bigr|_{k-m,\infty,\Omega}.$$

Notice that in the above, the constants

$$\max_{x\in\bar{\Omega}} \prod_{\theta\in[\Theta_{i-1}\setminus\tilde{\Theta},\theta_i]} \|x-\theta\|$$

were replaced by the possibly larger but far less complicated constant $(h_{\Omega,\Theta})^{k-m-j}$. This gives the first inequality with

$$C_{n,p,j,k,m} := (m+j)! \sum_{i=k-m-j}^{k} \binom{i-1}{m+j+i-k} \frac{1}{(k-i)!(k+1-i-n/p)_i}$$
$$= \frac{(k-1)!}{(k-m-j-1)!(1-n/p)} \, {}_2F_1\Bigl({-m-j,\, 1-n/p \atop 1-k} ; 1 \Bigr).$$

By the Chu–Vandermonde convolution identity:

$${}_2F_1\Bigl({-n, b \atop c} ; 1 \Bigr) = \frac{(c-b)_n}{(c)_n},$$

which is the special case $a = -n$ of equation (14) in [E53, p. 61], it follows that

$$C_{n,p,j,k,m} = \frac{1}{(1-n/p)_{k-m-j}}.$$

The second inequality, which is proved in [Wa97], follows from the pointwise estimate (3.4.4). □

By considering the special case of Taylor interpolation at a point by polynomials of degree $\le k$, we obtain the following estimate of the distance of smooth functions from Π_k.

Corollary. *Suppose that $\Omega \subset \mathbb{R}^n$ is a bounded, open, star-shaped region that has a Lipschitz boundary. Then for $p > n$ and $0 \le j \le k+1$,*

(4.3.3)
$$\operatorname{dist}_{|\cdot|_{j,p,\Omega}}(f, \Pi_k) := \inf_{g\in\Pi_k} \bigl| f - g \bigr|_{j,p,\Omega}$$
$$\le \frac{1}{(1-n/p)_{k+1-j}} (\operatorname{diam}\Omega)^{k+1-j} \bigl| f \bigr|_{k+1,p,\Omega}, \quad \forall f \in W_p^{k+1}(\Omega).$$

Note that

$$\frac{1}{(1-n/p)_{k+1-j}} \to \infty \qquad \text{as } p \to n^+.$$

That an inequality of the form of (4.3.3) holds for $j = 0$, where the constant $1/(1-n/p)_{k+1-j}$ is replaced by some unknown constant depending only on n, k, and

p, is the content of the paper by Dechevski and Quak [DQ90]. From this, they obtain the corresponding "improved" version of the Bramble–Hilbert lemma (see [BH70]).

4.4. A related result of Lai and Wang. The only related result in the literature is an L_p-bound for the error in Hakopian interpolation given by Lai and Wang [LW84]. In that paper they show the following.

THEOREM 4.4.1 ([LW84, Theorem 1]). *Let $|\alpha| \le k-n$. Then for any positive integer $\ell \le k+|\alpha|-n+1$, we have*

(4.4.2)

$$\begin{aligned} &D^\alpha(f-\mathcal{H}_\Theta^{(n-1)})(x) \\ &\quad=(|\alpha|+n-1)\sum_{\mu_1=1}^{|\alpha|+n}\sum_{i_1=1}^{n}(x-\theta_{|\alpha|+n-\mu_1+1})_{i_1}\sum_{\mu_2=1}^{\mu_1}\sum_{i_2=1}^{n}(x-\theta_{|\alpha|+n-\mu_2+2})_{i_2} \\ &\qquad\times\cdots\times\sum_{\mu_\ell=1}^{\mu_{\ell-1}}\sum_{i_\ell=1}^{n}(x-\theta_{|\alpha|+n-\mu_\ell+\ell})_{i_\ell}\int_{[\underbrace{x,\ldots,x}_{\mu_\ell},\theta_1,\ldots,\theta_{|\alpha|+n-\mu_\ell+\ell}]} D^{\alpha+\sum_{j=1}^{\ell}e^{i_j}}f \\ &\qquad-\sum_{j=|\alpha|+n-1+\ell}^{k-1}\ \sum_{|\gamma|=j-n+1} D^\alpha\omega_\gamma(x)\int_{[\theta_1,\ldots,\theta_j]} D^\gamma f. \end{aligned}$$

The above uses standard multiindex notation. The ith component of $x \in \mathbb{R}^n$ is x_i, and e^i is the ith unit vector in $\mathbb{R}^n$. To (4.4.2), Lai and Wang apply the integral form of Minkowski's inequality in the form

$$\left\| x \mapsto \int_{[\underbrace{x,\ldots,x}_{\mu},\theta_1,\ldots,\theta_{k+1-\mu}]} D^\beta f \right\|_{L_p(G)} \le C_2\, \|D^\beta f\|_{L_p(G)}, \quad \mu = 1,\ldots,|\alpha|+n, \tag{4.4.3}$$

to obtain the following.

THEOREM 4.4.4 ([LW84, Theorem 2]). *Let G be a convex set containing Θ with diameter h. If $p > n$, $|\alpha| \le k-n$, and $f \in W_p^{(k-n+1)}(G)$, then*

$$\|D^\alpha(f-\mathcal{H}_\Theta^{(n-1)}f)\|_{L_p(G)} \le C\, h^{k-n+1-|\alpha|} \max_{|\beta|=k-n+1} \|D^\beta f\|_{L_p(G)}, \tag{4.4.5}$$

where C is a constant independent of f.

Since $f \mapsto \max_{|\beta|=k+1-n} \|D^\beta f\|_{L_p(\Omega)}$ and $f \mapsto \big|f\big|_{k+1-n,p,\Omega}$ are equivalent seminorms, Theorem 4.4.4 follows from Theorem 4.3.1. Had Lai and Wang attempted to compute the C_2 of (4.4.3) using the multivariate form of Hardy's inequality, they would have obtained

$$C_2 \le \frac{1}{(\mu-1)!(\mu-n/p)_{k+1-\mu}}.$$

Thus their constant C in (4.4.5) would have the same qualitative behavior as our $C_{n,p,j,k,m}$ of (4.3.2), namely, that $C \to \infty$ as $p \to n^+$.

4.5. The behavior of $C_{n,p,j,k,m}$ as a function of its parameters. In [Wa97], it is shown that, in an appropriate sense, the constant $C_{n,p,j,k,m}$ of (4.3.2) is the best possible when $p = \infty$. The question then arises whether or not the overestimation committed in using the multivariate form of Hardy's inequality to obtain $C_{n,p,j,k,m}$ is significant for $p < \infty$. In particular, does the best possible constant C in the inequality

$$\left|f - \mathcal{H}_{\Theta}^{(m)} f\right|_{j,p,\Omega} \leq C\,(h_{\Omega,\Theta})^{k-m-j} \left|f\right|_{k-m,p,\Omega} \tag{4.5.1}$$

become unbounded as $p \to n^+$? In the univariate case, at least, the answer is *no*—the best possible constant in (4.5.1) does not become unbounded.

Before we show this, let us clarify a little the role that the condition $p > n$ plays in Theorems 4.3.1 and 4.4.4. The condition $p > n$ is necessary if these results are to be stated in terms of the Sobolev space $W_p^{(k-m)}(\Omega)$—in particular, so that $\mathcal{H}_{\Theta}^{(m)} f$ is defined for $f \in W_p^{(k-m)}(\Omega)$. However, it makes good sense to ask what the best constant C is for which (4.5.1) holds for all sufficiently smooth functions f—say, e.g., $f \in C^{k-m}(\bar{\Omega})$. The condition $p > n$ is again needed when we seek to apply the multivariate form of Hardy's inequality to the integral error formulas (4.2.2) and (4.4.2).

We now show that in the univariate case, i.e., when $n = 1$, there is a best possible constant C in (4.5.1) for all sufficiently smooth f which can be bounded independently of $1 \leq p \leq \infty$. The crucial step in the argument to follow is the use of the B-spline L_p-estimate that

$$\|M(\cdot|\Theta)\|_{L_p(\mathbb{R})} \leq \left(\frac{\#\Theta - 1}{\operatorname{diam}\Theta}\right)^{1-1/p} \tag{4.5.2}$$

when $\operatorname{diam}\Theta > 0$; see de Boor [B73].

In line with [Wa97], the univariate case of the map $\mathcal{H}_{\Theta}^{(m)}$, termed the *generalized Hermite map*, will be emphasized by writing it as $H_{\Theta}^{(m)}$. This map has the simple form

$$H_{\Theta}^{(m)} f = D^m(H_{\Theta} D^{-m} f),$$

where H_{Θ} is the *Hermite interpolator* at the points Θ and $D^{-m}f$ is any function for which $D^m(D^{-m}f) = f$.

THEOREM 4.5.3. *Let Θ be a k-sequence in the interval $[a\,.\,.\,b]$. If $1 \leq p, q \leq \infty$, $0 \leq j < k - m$, and $f \in C^{k-m}[a\,.\,.\,b]$, then*

$$\|D^j(f - H_{\Theta}^{(m)} f)\|_{L_p[a..b]} \leq \frac{(m+j)!}{(k-m-j)!}\,\frac{k^{1/q}}{k!}\,(b-a)^{k-m+\frac{1}{p}-\frac{1}{q}}\,\|D^{k-m}f\|_{L_q[a..b]}.$$

Proof. Fix $x \in [a\,.\,.\,b]$. For Θ a finite sequence in $\mathbb{R}$, let

$$\omega_{\Theta}(x) := \prod_{\theta\in\Theta}(x - \theta).$$

With this notation, replacing each occurrence in (4.2.2) of a linear functional of the form $f \mapsto \int_{\Theta} f$ by integration against a B-spline, we obtain that

$$\begin{aligned} &D^j(f - H_{\Theta}^{(m)} f)(x) \\ &\quad = (m+j)! \sum_{i=k-m-j}^{k} \sum_{\substack{\tilde{\Theta}\subset\Theta_{i-1} \\ \#\tilde{\Theta}=m+j+i-k}} \omega_{\Theta_{i-1}\backslash\tilde{\Theta}}(x)\,(x-\theta_i)\,\frac{1}{k!}\int D^{k-m} f\, M(\cdot|x,\Theta_i). \end{aligned}$$

By Hölder's inequality and (4.5.2), we have that

$$\left|\int D^{k-m} f\, M(\cdot|x,\Theta_i)\right| \le \left(\frac{k}{\operatorname{diam}[x,\Theta_i]}\right)^{1/q} \|D^{k-m}f\|_{L_q[a..b]}.$$

Since

$$\left|\frac{\omega_{\Theta_{i-1}\backslash\tilde{\Theta}}(x)\,(x-\theta_i)}{(\operatorname{diam}[x,\Theta_i])^{1/q}}\right| \le (b-a)^{k-m-1/q},$$

we obtain that

$$\begin{aligned}
&|D^j(f-H_\Theta^{(m)}f)(x)| \\
&\quad\le (m+j)! \sum_{i=k-m-j}^{k} \binom{i-1}{m+j+i-k} \frac{k^{1/q}}{k!}(b-a)^{k-m-1/q}\|D^{k-m}f\|_{L_q[a..b]} \\
&\quad= \frac{(m+j)!}{(k-m-j)!}\frac{k^{1/q}}{k!}(b-a)^{k-m-1/q}\|D^{k-m}f\|_{L_q[a..b]}.
\end{aligned}$$

Finally, take $\|\cdot\|_{L_q[a..b]}$ of both sides. □

To adapt this argument to the multivariate case, it is necessary to have the *simplex spline* analogue of the B-spline L_p-estimate (4.5.2). This is provided by Dahmen [D79], who shows that when $\operatorname{vol}_n(\operatorname{conv}\Theta) > 0$,

$$\|M(\cdot|\Theta)\|_{L_p(\mathbb{R}^n)} \le \frac{k!(k+1)!}{n!(n+1)!(n-k)!}\left(\frac{1}{\operatorname{vol}_n(\operatorname{conv}\Theta)}\right)^{1-1/p} \tag{4.5.4}$$

with $k+1 := \#\Theta$. Yet, with this in hand, it does not seem possible to apply the argument of Theorem 4.5.3 in any satifactory form.

Remark 4.5.5. Incidentally, the constant in (4.5.4) is not the best possible. Already, by using the fact that $\int M(\cdot|\Theta) = 1$ together with the case of (4.5.4) where $p=\infty$, we obtain

$$\|M(\cdot|\Theta)\|_{L_p(\mathbb{R}^n)} \le \left(\frac{k!(k+1)!}{n!(n+1)!(n-k)!}\frac{1}{\operatorname{vol}_n(\operatorname{conv}\Theta)}\right)^{1-1/p}.$$

In the univariate case this overestimates (4.5.2) by a factor of $((k+1)!/2)^{1-1/p}$.

The key step in proving (4.5.2) is the bound

$$M(\cdot|\Theta) \le \frac{k}{\operatorname{diam}\Theta}, \tag{4.5.6}$$

which follows from the partition of unity property of B-splines. Thus, a close examination of the simplex-spline analogue of the B-spline partition of unity, given recently by Dahmen, Micchelli, and Seidel [DMS92], should give tighter bounds than those of (4.5.4). However, we make no attempt here to give such an argument.

Remark 4.5.7. There are other integral error formulas for the scale of mean-value interpolations to which Theorem 3.4.1 can be applied to give L_p-bounds. These include those of Lai and Wang [LW86] (Kergin interpolation), Gao [Ga88], and Hakopian [BHS93, p. 200] (Hakopian interpolation). See [Wa97] for a discussion of the relative merits of each of these formulas.

5. Application: L_p-error bounds for multivariate Lagrange interpolation. In this section, we use Theorem 3.4.1 to obtain L_p-error bounds for *multivariate Lagrange interpolation schemes.*

5.1. Lagrange maps. A linear interpolation problem for which the space of interpolation conditions is spanned by *point evaluations* at Θ, a finite sequence in $\mathbb{R}^n$, is called a *Lagrange interpolation problem.* If P is the space of interpolants for such a problem and the problem is correct, then the associated linear projector, called the *Lagrange map*, will be denoted by $L_{P,\Theta}$. The *Lagrange form* of a Lagrange map is given by

$$L_{P,\Theta} f = \sum_{\theta \in \Theta} f(\theta) \ell_\theta. \tag{5.1.1}$$

Here (5.1.1) uniquely defines

$$\ell_\theta := \ell_{\theta,P,\Theta} \in P,$$

the *Lagrange function* for $\theta \in \Theta$. In other words, $(\delta_{[\theta]})_{\theta\in\Theta}$ is dual (biorthonormal) to $(\ell_\theta)_{\theta\in\Theta}$.

Lagrange maps into a space containing polynomials of degree k are frequently used to interpolate to scattered data; see, e.g., Alfeld [Al89]. Particular examples receiving much attention lately are maps where the interpolants include *radial basis functions* or *multivariate splines* and de Boor and Ron's *least solution* for the polynomial interpolation problem [BR90] (also see [BR92] for its generalization). In addition, there are, of course, the maps of Kergin and Hakopian.

For such maps, there is the *multipoint Taylor formula* for the error. This formula was initiated by the work of Ciarlet and Wagschal [CW71]; most of the relevant papers are in French, and it is not well known outside the area of finite elements. It is for these reasons and because our Theorem 3.4.1 implies L_p-estimates from the multipoint Taylor formula that we discuss the formula here.

5.2. The multipoint Taylor formula.

FORMULA 5.2.1 (multipoint Taylor formula; see [CR72]). *Let Θ be a finite sequence in $\mathbb{R}^n$, and let Ω be an open set in $\mathbb{R}^n$ for which $\bar\Omega$ is star-shaped with respect to Θ. If $L_{P,\Theta}$ is a Lagrange map with $\Pi_k(\mathbb{R}^n) \subset P \subset C^k(\bar\Omega)$, then for $f \in C^{k+1}(\bar\Omega)$, $q \in \Pi_k(\mathbb{R}^n)$, and $x \in \bar\Omega$, its error satisfies*

$$\big(q(D)(f - L_{P,\Theta} f)\big)(x) = -\sum_{\theta\in\Theta} \left(\int_{[\underbrace{x,\dots,x}_{k+1},\theta]} D^{k+1}_{\theta-x} f \right) (q(D)\ell_\theta)(x). \tag{5.2.2}$$

The term *multipoint Taylor formula* comes from the fact that

$$\theta \mapsto \int_{[\underbrace{x,\dots,x}_{k+1},\theta]} D^{k+1}_{\theta-x} f$$

is the error in *Taylor interpolation* of degree k at the point x, a special case of the error in Kergin interpolation. The proof of (5.2.2) further justifies the use of this term.

The region of integration in (5.2.2) consists of line segments from x to $\theta \in \Theta$ (see Figure 5.1).

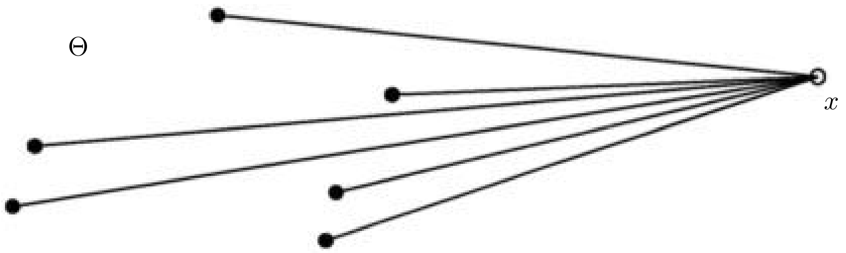


FIG. 5.1. *The region of integration in* (5.2.2) *for* Θ *consisting of six points.*

From the multipoint Taylor formula, Arcangeli and Gout [AG76] obtain L_p-bounds for the error in a Lagrange map. These bounds are precisely those obtained by applying Theorem 3.4.1 to (5.2.2). The crucial step in the argument presented in [AG76, Proposition 1-1] is the use of the multivariate form of Hardy's inequality for the map

$$x \mapsto H_{k+1,[v]}f(x) := \int_{[\underbrace{x,\ldots,x}_{k+1},v]} f. \tag{5.2.3}$$

This inequality is not explicitly stated, though the proof of their (weaker) Proposition 1–1 would imply it.

Remark 5.2.4. The key step in the proof of Proposition 1-1 in [AG76] is an application of Hölder's inequality to the splitting

$$\int_{[\underbrace{x,\ldots,x}_{k+1},v]} f = \frac{1}{k!}\int_0^1 (1-t)^{-1/q-\varepsilon}\left((1-t)^{k+1/q-\varepsilon}f(x+t(v-x))\right)dt,$$

where $\varepsilon := (k+1-n/p)/q$ and $1/p+1/q=1$, as opposed to our use of the integral form of Minkowski's inequality.

Having identified the precise role of the multivariate form of Hardy's inequality in [AG76], it is possible to use it to run through Arcangeli and Gout's calculation for a much more general class of norms, including those most often used in numerical analysis. The resulting bounds, given below, have smaller (and simpler) constants than those one might hope to obtain by applying the inequalities for similar norms to the results of [AG76].

For the remainder of this section, Ω will denote a bounded open set in $\mathbb{R}^n$ with a Lipschitz boundary, and Θ will denote a finite sequence in $\mathbb{R}^n$. Recall that

$$h_{\Omega,\Theta} = \sup_{x\in\Omega}\sup_{\theta\in\Theta}\|x-\theta\| \le \operatorname{diam}\Omega.$$

COROLLARY 5.2.5. *Suppose that $\bar{\Omega}$ is star-shaped with respect to Θ and that $L_{P,\Theta}$ is a Lagrange map with $\Pi_k(\mathbb{R}^n) \subset P \subset C^k(\Omega)$. If $k+1-n/p>0$, $g\in\Pi_k$, and $f \in W_p^{(k+1)}(\Omega)$, then*

(5.2.6)

$$\|g(D)(f-L_{P,\Theta}f)\|_{L_p(\Omega)} \le \frac{1}{k!(k+1-n/p)}\left(\sum_{\theta\in\Theta}\|g(D)\ell_\theta\|_{L_\infty(\Omega)}\right)\bigl|f\bigr|_{k+1,p,\Omega}(h_{\Omega,\Theta})^{k+1},$$

and so, in particular,

$$(5.2.7) \qquad |f - L_{P,\Theta}f|_{p,\Omega} \le \frac{1}{k!(k+1-n/p)} \Bigl(\sum_{\theta\in\Theta} |\ell_\theta|_{\infty,\Omega} \Bigr) \bigl|f\bigr|_{k+1,p,\Omega} \, (h_{\Omega,\Theta})^{k+1},$$

where $|\cdot|_{p,\Omega}$ is any seminorm on $W_p^k(\Omega)$ of the form

$$|f|_{p,\Omega} := \| \, (\|g_i(D)f\|_{L_p(\Omega)})_{i=1}^m \, \|_{\mathbb{R}^m},$$

where the $g_i \in \Pi_k(\mathbb{R}^n)$'s are fixed and $\|\cdot\|_{\mathbb{R}^m}$ is any monotone norm on $\mathbb{R}^m$.

Proof. By Sobolev's embedding theorem, the condition $k+1-n/p > 0$ implies that

$$W_p^{(k+1)}(\Omega) \subset C(\bar\Omega),$$

and so the Lagrange map $L_{P,\Theta}$ is well defined. As in Remark 4.2.3, (5.2.2) can be extended to $f \in W_p^{(k+1)}(\Omega)$. Fix $f \in W_p^{(k+1)}(\Omega)$ and $x \in \Omega$. Let $h := h_{\Omega,\Theta}$. By (1.4.2),

$$|D^{k+1}_{\theta-x}f| \le |D^{k+1}f| \, \|\theta - x\|^{k+1} \le |D^{k+1}f| \, h^{k+1}$$

in $L_p(\Omega)$. Thus, from (5.2.2), it follows that for a.e. $x \in \Omega$,

$$|(g(D)(f - L_{P,\Theta}f))(x)| \le \sum_{\theta\in\Theta} \Bigl(\int_{[\underbrace{x,\ldots,x}_{k+1},\theta]} |D^{k+1}f| \Bigr) \|g(D)\ell_\theta\|_{L_\infty(\Omega)} \, h^{k+1}.$$

To this, the condition $k+1-n/p > 0$ allows us to apply the multivariate form of Hardy's inequality to obtain (5.2.6). □

In [AG76, Theorem 1-1], (5.2.7) is proved when $|\cdot|_{p,\Omega}$ is of the form $|\cdot|_{i,p,\Omega}$ for some $0 \le i \le k$, with $h_{\Omega,\Theta}$ replaced by $\operatorname{diam}\Omega$. In that paper, some bounds on the size of the Lagrange functions ℓ_θ together with relevant applications are given. One application is bounding the error in a *finite element scheme*; also see Ciarlet [Ci78, p. 128]. Another, of interest to approximation theorists, is to estimate the distance of smooth functions from $\Pi_k(\mathbb{R}^n)$ and to give the corresponding *constructive* version of the Bramble–Hilbert lemma; see [BH70].

The condition in Corollary 5.2.5 that $k+1-n/p > 0$ plays an analogous role to the condition in Theorem 4.3.1 that $n > p$. Namely, it is required so that the results can be stated in terms of Sobolev spaces and to apply the multivariate form of Hardy's inequality. Additionally, by Theorem 4.5.3, the unboundedness of the constant in (5.2.7) as $k+1-n/p \to 0^+$ is not a true reflection of the behavior of the error in the invariate case.

With the multivariate form of Hardy's inequality in hand, it is also possible to obtain pointwise error bounds for Lagrange maps.

Corollary 5.2.8. *Suppose that $\bar\Omega$ is star-shaped with respect to Θ and that $L_{P,\Theta}$ is a Lagrange map with $\Pi_k(\mathbb{R}^n) \subset P \subset C^k(\Omega)$. With $f \in W_\infty^{(k+1)} \subset C(\bar\Omega)$ and $x \in \bar\Omega$, we have the (coordinate-independent) pointwise error bound*

$$(5.2.9) \qquad |f(x) - L_{P,\Theta}f(x)| \le \frac{1}{(k+1)!} \bigl|f\bigr|_{k+1,\infty,\Omega} \sum_{\theta\in\Theta} \|\theta - x\|^{k+1} |\ell_\theta(x)|$$

and the (coordinate-dependent) pointwise error bound

$$(5.2.10) \qquad |f(x) - L_{P,\Theta}f(x)| \le \sum_{\theta\in\Theta} \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|D^\alpha f\|_{L_\infty(\Omega)} \, |(\theta - x)^\alpha \ell_\theta(x)|.$$

Proof. The proof runs along the same lines as that for Corollary 5.2.5, except that for (5.2.10), we first expand $D_{\theta-x}^{k+1}f$ as

$$D_{\theta-x}^{k+1}f = \sum_{|\alpha|=k+1} \frac{(k+1)!}{\alpha!}(\theta-x)^\alpha D^\alpha f$$

by using the multinomial identity. □

Neither of (5.2.9) or (5.2.10) occurs in the literature. For $f \in C^{k+1}(\Omega)$, they can be obtained more simply by applying the mean-value theorem, as given by Properties 2.1.3(c), to the integrals occurring in (5.2.2).

Remark 5.2.10. The results of [AG76] have been extended in the following ways. In [Go77], Gout treats the error in certain forms of *Hermite interpolation*—that is where, in addition to function values, certain derivatives are matched at the points in Θ. In [AS84], Arcangeli and Sanchez bound the error in a Lagrange map for functions from *fractional-order* Sobolev spaces.

5.3. The error formula of Sauer and Xu. There is another error formula, one for the error in a Lagrange map with range (interpolants) $\Pi_k(\mathbb{R}^n)$, that has been given recently by Sauer and Xu; see [SX95].

Sauer and Xu order the $\dim \Pi_k(\mathbb{R}^n)$ points in Θ so that each Lagrange interpolation problem with points Θ^j (by definition, the initial segment of Θ consisting of the first $\dim \Pi_j(\mathbb{R}^n)$ terms) and interpolants $\Pi_j(\mathbb{R}^n)$ is correct for $j = 0, \ldots, k$. They consider the collection $\boldsymbol{\Psi}$ of all $(k+1)$-sequences $\Psi = [\psi_0, \ldots, \psi_k]$, which they call *paths*, with $\psi_j \in \Theta^j \backslash \Theta^{j-1}$ for all j. Given this notation, Sauer and Xu state their result in the following form.

THEOREM 5.3.1 ([SX95, Theorem 3.6]). *Suppose that $L_{P,\Theta} := L_{\Pi_k(\mathbb{R}^n),\Theta}$ is a Lagrange map and that $f \in C^{k+1}(\mathbb{R}^n)$. Then*

$$L_{P,\Theta}f(x) - f(x) = \sum_{\Psi \in \boldsymbol{\Psi}} p_\Psi(x) \int_{[x,\Psi]} D_{x-\psi_k} D_{\psi_k-\psi_{k-1}} \cdots D_{\psi_2-\psi_1} D_{\psi_1-\psi_0} f, \tag{5.3.2}$$

where $p_\Psi \in \Pi_k(\mathbb{R}^n)$ is given by

$$p_\Psi(x) := (k+1)!\, \ell_{\psi_k,\Pi_k(\mathbb{R}^n),\Theta}(x) \prod_{i=1}^{k} \ell_{\psi_i,\Pi_i(\mathbb{R}^n),\Theta^i}(\psi_{i+1}).$$

The region of integration in each term of (5.3.2) is the convex hull of x and Ψ (see Figure 5.2).

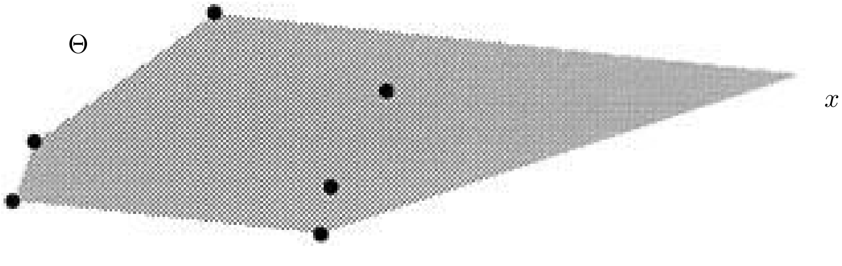


FIG. 5.2. *The region of integration in* (5.3.2) *for* Θ *consisting of six points.*

From (5.3.2), the following pointwise estimate is obtained.

COROLLARY 5.3.3 ([SX95, Corollary 3.11]). *Suppose, in addition to the hypotheses of Theorem 5.3.1, that $\bar{\Omega}$ is star-shaped with respect to Θ. Then for all $x \in \bar{\Omega}$,*

$$\begin{aligned} &|f(x) - L_{P,\Theta}f(x)| \\ &\qquad \le \frac{1}{(k+1)!} \sum_{\Psi \in \mathbf{\Psi}} \|D_{x-\psi_k} D_{\psi_k - \psi_{k-1}} \cdots D_{\psi_2 - \psi_1} D_{\psi_1 - \psi_0} f\|_{L_\infty(\Omega)} |p_\Psi(x)|. \end{aligned} \tag{5.3.4}$$

The bound (5.3.4) is of a form similar to those of (5.2.9) and (5.2.10). For a more direct comparison, we obtain from (5.2.2) the bound

$$|f(x) - L_{P,\Theta}f(x)| \le \frac{1}{(k+1)!} \sum_{\theta \in \Theta} \|D_{\theta - x}^{k+1} f\|_{L_\infty(\Omega)} |\ell_\theta(x)|. \tag{5.3.5}$$

This last bound has $\#\Theta = \sum_{j=0}^{k} \#\Theta^j$ terms, as opposed to $\#\mathbf{\Psi} = \prod_{j=0}^{k} \#\Theta^j$ for (5.3.4), and requires no ordering of Θ. For the purposes of comparison, in the bivariate case, i.e., when $n = 2$, we have that $\#\Theta = (k+2)(k+1)/2$, while $\#\mathbf{\Psi} = (k+1)!$. In addition, bounds analogous to (5.3.5) can be obtained from (5.2.2) for the derivatives of the error in $L_{P,\Theta}$.

To obtain L_p-bounds from (5.3.2), it is necessary to bound

$$x \mapsto L_{1,\Psi} f(x) := \int_{[x,\Psi]} f \tag{5.3.6}$$

in terms of $\|f\|_{L_p(\Omega)}$. This can be done by using the multivariate form of Hardy's inequality. Thus, we have the following instance of Theorem 3.4.1.

COROLLARY 5.3.7. *Suppose the hypotheses of Corollary* 5.3.3. *If $1 - n/p > 0$, then*

$$\|f - L_{P,\Theta} f\|_{L_p(\Omega)} \le \frac{1}{(1-n/p)_{k+1}} \left(\sum_{\Psi \in \mathbf{\Psi}} \|p_\Psi\|_{L_\infty(\Omega)} \right) |f|_{k+1,p,\Omega} (h_{\Omega,\Theta})^{k+1}.$$

The condition $1 - n/p > 0$ is needed so that the multivariate form of Hardy's inequality can be applied to (5.3.6). By comparison, to obtain (5.2.7) from (5.2.3), only the weaker condition that $k + 1 - n/p > 0$ was needed.

6. Other error bounds.

6.1. Discussion. Most of the integral error formulas for Lagrange maps given in the literature, including those of section 5, can be obtained from

$$f(x) - L_{P,\Theta} f(x) = \sum_{\theta \in \Theta} \left(\int_{[x]} f - \int_{[\theta]} f \right) \ell_\theta(x),$$

which is valid whenever P contains the constants, by appropriately using the identity

$$\int_{[\Theta,v]} f - \int_{[\Theta,w]} f = \int_{[\Theta,v,w]} D_{v-w} f \tag{6.1.1}$$

and integration by parts.

For example, in Gregory [Gr75] integration by parts is used to give a *Taylor-type* expansion for f. From this is obtained an integral error formula for *linear interpolation* on a triangle, i.e., when Θ consists of three affinely independent points in $\mathrm{I\!R}^2$, and the interpolants are the linear polynomials $P := \Pi_1(\mathrm{I\!R}^2)$. Such an argument is frequently referred to as a *Sard kernel theory* argument, as developed by Sard [Sa63]. The resulting formula is complicated—it has four line integrals and five area integrals. Another example is given by Hakopian [H82], who uses (6.1.1) to obtain an integral error formula for *tensor-product* Lagrange interpolation.

In view of their derivations, all of these integral error formulas involve terms which consist of a function (obtained appropriately from the Lagrange functions) multiplied by the integral of some derivative against a simplex spline. Thus it is possible to apply the multivariate form of Hardy's inequality to all such formulas (and those likely to be obtained in the future) to obtain L_p-bounds—with the caution that, as pointed out for the examples in sections 4 and 5, for small p, this may not accurately reflect the behavior of the error.

Exactly how to use (6.1.1) and integration by parts to obtain the best possible error formula for a given purpose is far from clear. In a future paper, the author considers the simplest case, that of linear interpolation on a triangle. There the formulas of Ciarlet and Wagschal [CW71], Gregory [Gr75], Sauer and Xu [SX95], and others are discussed.

Acknowledgments. The author would like to thank the referees, Carl de Boor, Shaun Cooper, and Geoff Pritchard for their considered comments and help with this paper.

REFERENCES

[Ad75] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[Al89] P. Alfeld, *Scattered data interpolation in three or more variables*, in Mathematical Methods in Computer Aided Geometric Design, T. Lyche and L. Schumaker, eds., Academic Press, New York, 1989, pp. 1–33.

[AG76] R. Arcangeli and J. L. Gout, *Sur l'evaluation de l'erreur d'interpolation de Lagrange dans un ouvert de* $\mathrm{I\!R}^n$, Rev. Française Automat. Inform. Rech. Opér. Anal. Numer., 10 (1976), pp. 5–27.

[AS84] R. Arcangeli and A. M. Sanchez, *Estimations des erreurs de meilleure approximation polynomiale et d'interpolation de Lagrange dans les espaces de Sobolev d'ordre non entier*, Numer. Math., 45 (1984), pp. 301–321.

[BHS93] B. D. Bojanov, H. A. Hakopian, and A. A. Sahakian, *Spline functions and multivariate interpolations*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

[B73] C. de Boor, *The quasi-interpolant as a tool in elementary spline theory*, in Approximation Theory, G. G. Lorentz, H. Berens, E. W. Cheney, and L. L. Schumaker, eds., Academic Press, New York, 1973, pp. 269–276.

[BR90] C. de Boor and A. Ron, *On multivariate polynomial interpolation*, Constr. Approx., 6 (1990), pp. 287–302.

[BR92] C. de Boor and A. Ron, *The least solution for the polynomial interpolation problem*, Math. Z., 210 (1992), pp. 347–378.

[BH70] J. H. Bramble and S. R. Hilbert, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[CD90] W. Chen and Z. Ditzian, *Mixed and directional derivatives*, Proc. Amer. Math. Soc., 108 (1990), pp. 177–185.

[DQ90] L. T. Dechevski and E. Quak, *On the Bramble–Hilbert lemma*, Numer. Funct. Anal. Optim., 11 (1990), pp. 485–495.

[Ci78] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[CR72] P. G. Ciarlet and P. A. Raviart, *General Lagrange and Hermite interpolation in* $\mathbb{R}^N$ *with applications to finite element methods*, Arch. Rational Mech. Anal., 46 (1972), pp. 177–199.

[CW71] P. G. Ciarlet and C. Wagschal, *Multipoint Taylor formulas and applications to the finite element method*, Numer. Math., 17 (1971), pp. 84–100.

[D79] W. Dahmen, *Multivariate B-splines: Recurrence relations and linear combinations of truncated powers*, in Multivariate Approximation Theory, W. Schempp and K. Zeller, eds., Birkhäuser, Basel, 1979, pp. 64–82.

[DMS92] W. Dahmen, C. A. Micchelli, and H.-P. Seidel, *Blossoming begets B-spline bases built better by B-patches*, Math. Comp., 59 (1992), pp. 97–115.

[E53] A. Erdélyi, *Higher transcendental functions*, Vol. I, McGraw–Hill, New York, 1953.

[FMP91] A. M. Fink, D. S. Mitrinović, and J. E. Pečarić, *Inequalities involving functions and their integrals and derivatives*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991.

[Fo84] G. B. Folland, *Real Analysis, Modern Techniques and Their Applications*, John Wiley, New York, 1984.

[Ga88] J. B. Gao, *Multivariate quasi-Newton interpolation*, J. Math. Res. Exposition, 8(3) (1988), pp. 447–453.

[Ge1869] A. Genocchi, *Relation entre la différence et la dérivée d'un même ordre quelconque*, Arch. Math. Phys. (I), 49 (1869), pp. 342–345.

[Go77] J. L. Gout, *Estimation de l'erreur d'interpolation d'Hermite dans* $\mathbb{R}^n$, Numer. Math., 28 (1977), pp. 407–429.

[Gr75] J. A. Gregory, *Error bounds for linear interpolation on triangles*, in Mathematics of Finite Elements and Applications, J. Whiteman, ed, Academic Press, London, 1975, pp. 163–170.

[H82] H. Hakopian, *Integral remainder formula of the tensor product interpolation*, Bull. Polish Acad. Sci. Math., 31 (1982), pp. 267–272.

[Ha28] G. H. Hardy, *Notes on some points in the integral calculus* LXIV, Messenger of Math., 57 (1928), pp. 12–16.

[HLP67] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge University Press, Cambridge, 1967.

[HAR96] L. A. Harris, *Bernstein's polynomial inequalities and functional analysis*, Irish Math. Soc. Bull., 36 (1996), pp. 19–33.

[LW84] M. Lai and X. Wang, *A note to the remainder of a multivariate interpolation polynomial*, Approx. Theory Appl., 1 (1984), pp. 57–63.

[LW86] M. Lai and X. Wang, *On multivariate Newtonian interpolation*, Sci. Sinica Ser. A, 29 (1986), pp. 23–32.

[M79] C. A. Micchelli, *On a numerically efficient method for computing multivariate B-splines*, in Multivariate Approximation Theory, W. Schempp and K. Zeller, eds., Birkhäuser, Basel, 1979, pp. 211–248.

[M80] C. A. Micchelli, *A constructive approach to Kergin interpolation in* $\mathbb{R}^k$*: Multivariate B-splines and Lagrange interpolation*, Rocky Mountain J. Math., 10 (1980), pp. 485–497.

[Ru87] W. Rudin, *Real and Complex analysis*, McGraw–Hill, New York, 1987.

[Sa63] A. Sard, *Linear Approximation*, AMS, Providence, RI, 1963.

[SX95] T. Sauer and Yuan Xu, *On multivariate Lagrange interpolation*, Math. Comp., 64 (1995), pp. 1147–1170.

[Wa97] S. Waldron, *Integral error formulæ for the scale of mean value interpolations which includes Kergin and Hakopian interpolation*, Numer. Math., to appear, 1997.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 259–269, March 1997

BLOWUP IN REACTION-DIFFUSION SYSTEMS WITH DISSIPATION OF MASS*

MICHEL PIERRE† AND DIDIER SCHMITT†

Abstract. We prove blowup in finite time of the solutions to some reaction-diffusion systems that preserve nonnegativity and for which the total mass of the components is uniformly bounded. (These are natural properties in applications.) This is done by presenting explicit counterexamples constructed with the help of formal computation software. Several partial results of global existence had been obtained previously in the literature. Our counterexamples explain a posteriori why extra conditions are needed. Negative results are also provided as a by-product for linear parabolic equations in nondivergence form and with discontinuous coefficients and for nonlinear Hamilton–Jacobi evolution equations.

Key words. parabolic system, reaction-diffusion, global existence, blowup, parabolic equation in nondivergence form, Hamilton–Jacobi equation

AMS subject classifications. 35K10, 35K45, 35K57

PII. S0036141095295437

1. Introduction.

We are mainly interested in global existence in time or blowup in finite time of solutions to reaction-diffusion systems of the form

$$\frac{\partial u}{\partial t} - d_1 \Delta u = f(u,v) \quad \text{on } (0,\infty)\times\Omega, \tag{1}$$

$$\frac{\partial v}{\partial t} - d_2 \Delta v = g(u,v) \quad \text{on } (0,\infty)\times\Omega \tag{2}$$

for which the following two main properties hold:

- the positivity of the solutions is preserved with time, which is equivalent to

$$\forall u, v \geq 0, \quad f(0,v) \geq 0, \qquad g(u,0) \geq 0; \tag{3}$$

- the total mass of the components u and v is nonincreasing with time, which is essentially ensured by the structure condition

$$f + g \leq 0. \tag{4}$$

Here f and g are regular functions from $[0,\infty[^2$ into $I\!R, d_1$ and d_2 are positive constants, and Ω is a smooth bounded open subset of $I\!R^N$. As usual, "good" boundary conditions should be prescribed for u and v, for instance,

$$u = 0, \qquad v = 0 \quad \text{on } \partial\Omega, \tag{5}$$

as well as initial conditions,

$$u(0,.) = u_0 \geq 0, \qquad v(0,.) = v_0 \geq 0. \tag{6}$$

* Received by the editors November 28, 1995; accepted for publication January 15, 1996.
http://www.siam.org/journals/sima/28-2/29543.html

† CNRS UMR 9973, INRIA-Lorraine Projet NUMATH, Laboratoire de Mathématiques, Université Henri Poincaré–Nancy I, B.P. 239, 54506 Vandoeuvre-lès-Nancy cedex, France (pierre@loria.fr, dschmitt@iecn.u-nancy.fr).

The main consequence of properties (3) and (4) is that the solutions u and v satisfy an a priori L^1-estimate uniformly in time. Indeed, integrating the sum of (1) and (2) and taking (5) into account leads to

$$\frac{\partial}{\partial t}\int_\Omega u(t)+v(t)\le\int_\Omega f+g\le 0,$$

$$\int_\Omega u(t)+v(t)\le\int_\Omega u_0+v_0. \tag{7}$$

Since u and v are nonnegative, this is a uniform estimate of their L^1-norms. It is well known that local existence of nonnegative solutions holds for the system (1), (2), (5), (6) when $u_0, v_0\in L^\infty(\Omega)$. Moreover, existence is global as soon as $u(t)$ and $v(t)$ satisfy an a priori L^∞-estimate uniformly in time.

Here the a priori estimate is only in $L^1(\Omega)$. Much work has been done to analyze how this L^1-estimate or, more generally, structure conditions like (4) help to provide global existence.

Note, for instance, that if $d_1=d_2$, summing (1) and (2) leads—thanks to (4)—to

$$\frac{\partial}{\partial t}(u+v)-d_1\Delta(u+v)\le 0,$$

and by the maximum principle,

$$||(u+v)(t)||_\infty\le||u_0+v_0||_\infty,$$

so global existence holds.

Note also that properties (3) and (4) imply global existence (for nonnegative data) for the associated ordinary differential system

$$\dot u=f(u,v),\qquad \dot v=g(u,v).$$

For the complete system with different diffusion coefficients $d_1\ne d_2$, the question is considerably more delicate. One of the main result is that, in general, if one of u or v is a priori bounded, then so is the other. This is the case if, for instance, we have the extra information

$$f\le 0. \tag{8}$$

Obviously, by the maximum principle, this implies $||u(t)||_\infty\le||u_0||_\infty$. If g is of at most polynomial growth, then global existence can be proved [2], [4], [6], [7], [8].

However, it was still an open problem to decide what happens for systems without any a priori L^∞-bound on either u or v and whether extra conditions must be added to (3) and (4). This is precisely the goal of this paper, where we prove that blowup in L^∞ may occur in finite time for these systems.

In order to better understand the question, let us discuss some explicit examples of "systems" that naturally appear in this class. We have, for instance, the following:

$$u_t-d_1\Delta u=\lambda u^3v^2-u^2v^3, \tag{9}$$

$$v_t-d_2\Delta v=-u^2v^3+u^3v^2. \tag{10}$$

Here $f+g=(\lambda-1)u^3v^2 \leq 0$ if $\lambda \in [0,1]$. If $d_1=d_2$ or, more generally, if d_1 is close to d_2, then the system (9), (10), (5), (6) has a global solution. (See the remark above when $d_1=d_2$ and see [1] and [10] in general.)

Now if d_1 and d_2 are distinct and not close to each other, the question is more serious. If $\lambda=0$, we are in the situation described above where (8) also holds. Then u is a priori uniformly bounded and global existence can be proved. The case when λ is small enough can also be taken care of similarly [10]. Now the question is open when $\lambda \in \,]0,1]$. Note that here we even have

$$f+\lambda g=(\lambda-1)u^2v^3 \leq 0, \tag{11}$$

so if $\lambda \in [0,1[$, inequalities (4) and (11) are linearly independent. This implies, for instance, an a priori L^1-bound on the nonlinear terms.

We will see here that similar systems with these properties may present blowup in L^∞.

Let us also discuss another example, which is specifically studied in [3], [9], and [10]:

$$\begin{aligned} u_t-d_1u_{xx}&=-c(x)u^\alpha v^\beta,\\ v_t-d_2v_{xx}&=c(x)u^\alpha v^\beta, \end{aligned}$$

where $c:(-1,1)\to I\!R$ is given. Here $\Omega=(-1,1)$ and f and g also depend on the space variable x and satisfy $f+g=0$.

If $c\equiv 1$, we are in the situation of (8) and global existence follows. The same holds if $c(.)$ is of constant sign. Now the situation is quite different if $c(.)$ changes sign. The following specific case is analyzed in [9]:

$$c(x)>0 \quad \text{on } (0,1), \qquad c(0)=0, \qquad c(x)<0 \quad \text{on } (-1,0).$$

It can be shown that u and v are uniformly bounded in $L^\infty_{\text{loc}}([0,\infty)\times(0,1])$ and $L^\infty_{\text{loc}}([0,\infty)\times[-1,0))$. Therefore, blowup can only occur at $x=0$ and, if so, will occur for u and v at the same time. If c(.) vanishes fast enough at 0, then no blowup occurs (see [9]). The question is still open for a general $c(.)$. However, here we present similar examples where blowup in L^∞ does happen.

It is interesting to remark that in all of our examples, although solutions blow up in $L^\infty(\Omega)$ for some T, they can be extended across T as global "weak" solutions, for instance, in the sense of distributions.

Finally, let us mention that the blowup examples described in this paper provide interesting by-products for two questions of independent interest related to the following equations:

$$\begin{cases} u_t-a(x,t)\Delta u=f & \text{on } (0,T)\times\Omega,\\ u_{|\partial\Omega}=0, \qquad u(0,.)=0, \end{cases} \tag{12}$$

where

$$0<d_1\leq a(x,t)\leq d_2, \tag{13}$$

and

$$\begin{cases} u_t-\max(d_1\Delta u, d_2\Delta u)=f & \text{on } (0,T)\times\Omega,\\ u_{|\partial\Omega}=0, \qquad u(0,.)=0. \end{cases} \tag{14}$$

We prove that problem (14) is ill-posed for $f \in L^p(Q_T)$ when p is close to 1 (although it is well-posed if $p \geq 2$). For problem (12), we prove there is no estimate of the form

$$||u(T)||_{L^1(\Omega)} \leq C||f||_{L^p(Q_T)} \tag{15}$$

with a constant C depending only on d_1, d_2, T, and Ω when p is close to 1. Here also, such estimates are valid if $p \geq 2$. Note also that (15) holds when $a(.,.)$ is continuous with C depending on its modulus of continuity. Indeed, we have an L^p-theory, and u_t and Δu are bounded in L^p. Obviously, the question here concerns a subclass of parabolic operators of nondivergence form with discontinuous coefficients.

2. The main results. We denote by B the Euclidian unit ball in $I\!R^N$, $Q_T = (0,T) \times B$, and $\Sigma_T = (0,T) \times \partial B$.

THEOREM 2.1. *There exist $f, g \in C^\infty([0,\infty)^2, I\!R)$, $d_1, d_2 > 0$, $T > 0$, u_0, $v_0 \in C^\infty(\overline{B})$, $u_0 \geq 0$, $v_0 \geq 0$, $\alpha_1, \alpha_2 \in C^\infty([0,T])$, $\lambda \in]0,1[$, and $u, v \geq 0$, classical solutions of*

$$\frac{\partial u}{\partial t} - d_1 \Delta u = f(u,v) \quad \textit{on } Q_T, \tag{16}$$

$$\frac{\partial v}{\partial t} - d_2 \Delta v = g(u,v) \quad \textit{on } Q_T, \tag{17}$$

$$u(t,x) = \alpha_1(t), \qquad v(t,x) = \alpha_2(t) \quad \textit{on } \Sigma_T, \tag{18}$$

$$u(0,x) = u_0(x), \qquad v(0,x) = v_0(x) \quad \textit{on } B \tag{19}$$

such that

$$f + g \leq 0, \qquad f + \lambda g \leq 0, \tag{20}$$

$$\exists k > 0,\, p \geq 1, \quad \forall u, v \geq 0, \qquad |f(u,v)| + |g(u,v)| \leq k(u^p + v^p + 1), \tag{21}$$

$$\forall r, s \geq 0, \quad f(0,s) \geq 0, \qquad g(r,0) \geq 0, \tag{22}$$

and

$$\lim_{t \uparrow T} ||u(t)||_{L^\infty(B)} = \lim_{t \uparrow T} ||v(t)||_{L^\infty(B)} = +\infty. \tag{23}$$

THEOREM 2.2. *There exist $\alpha, \beta > 1$, $d_1, d_2 > 0$, $T > 0, u_0$, $v_0 \in C^\infty(\overline{B})$, $u_0 \geq 0$, $v_0 \geq 0$, $\alpha_1, \alpha_2 \in C^\infty([0,T])$, $c_1, c_2 \in C^k(\overline{Q}_T)$ with $k \geq 0$, and $u, v \geq 0$, classical solutions of*

$$\frac{\partial u}{\partial t} - d_1 \Delta u = c_1(t,x) u^\alpha v^\beta \quad \textit{on } Q_T, \tag{24}$$

$$\frac{\partial v}{\partial t} - d_2 \Delta v = c_2(t,x) u^\alpha v^\beta \quad \textit{on } Q_T, \tag{25}$$

$$u(t,x) = \alpha_1(t), \qquad v(t,x) = \alpha_2(t) \quad \textit{on } \Sigma_T, \tag{26}$$

$$u(0,x) = u_0(x), \qquad v(0,x) = v_0(x) \quad \textit{on } \Omega \tag{27}$$

such that

$$c_1(x,t) + c_2(x,t) \leq 0 \quad \textit{on } Q_T \tag{28}$$

and

$$\lim_{t\uparrow T} ||u(t)||_{L^\infty(B)} = \lim_{t\uparrow T} ||v(t)||_{L^\infty(B)} = +\infty. \tag{29}$$

Remark 2.1. In both theorems, u and v satisfy

$$\frac{\partial u}{\partial t} - d_1\Delta u + \frac{\partial v}{\partial t} - d_2\Delta v \leq 0. \tag{30}$$

Together with the boundary conditions on u and v, this implies, in particular, that the L^1-norms of $u(t)$ and $v(t)$ are bounded on $(0,T)$. Actually, as will appear clearly in the examples in the next section, in both cases, there exists $p^* \in (1,\infty)$ such that

$$\forall p < p^*, \quad \sup_{t\in(0,T)} (||u(t)||_{L^p(B)} + ||v(t)||_{L^p(B)}) < \infty, \tag{31}$$

$$\forall p \geq p^*, \quad \lim_{t\uparrow T} ||u(t)||_{L^p(B)} = \lim_{t\uparrow T} ||v(t)||_{L^p(B)} = +\infty. \tag{32}$$

The proofs of both theorems will be obtained by presenting explicit solutions u and v satisfying the fundamental inequality (30). Many consequences may be derived from this together with the nonnegativity of u and v. We have already mentioned the uniform L^1-bound on u and v. It also implies (see [2] and [6]) that for all $p \in (1,\infty)$, there exists $C = C(p, T, \Omega, \alpha_1, \alpha_2, ||u_0||_\infty, ||v_0||_\infty)$ such that

$$\forall t \in (0,T), \quad ||u||_{L^p(Q_T)} \leq C||v||_{L^p(Q_T)}, \qquad ||v||_{L^p(Q_T)} \leq C||u||_{L^p(Q_T)}. \tag{33}$$

Consequently, u and v can only blow up at the same time t when (30) holds.

Remark 2.2. Another interesting consequence of (30) may be obtained by setting

$$w(t,x) := u(t,x) + v(t,x), \qquad a(t,x) = \frac{d_1u(t,x) + d_2v(t,x)}{u(t,x) + v(t,x)}. \tag{34}$$

Then (30) can be rewritten as

$$\frac{\partial w}{\partial t} - \Delta(aw) \leq 0. \tag{35}$$

Here, thanks to the positivity of u and v, we have the a priori estimate

$$0 < \min(d_1, d_2) \leq a(t,x) \leq \max(d_1, d_2). \tag{36}$$

A natural question is whether the parabolic inequality (35) implies the existence of a constant C depending only on d_1, d_2, T, α_1, and α_2 such that for p large,

$$||w||_{L^p(Q_T)} \leq C(||w_0||_\infty + 1). \tag{37}$$

By duality, this is equivalent (see the next section) to the existence of a similar constant C such that

$$||z(T)||_{L^1(Q_T)} \leq C||\theta||_{L^q(Q_T)}, \tag{38}$$

where $1/q + 1/p = 1$ and z is solution of the dual problem

$$z_t - b\Delta z = \theta \quad \text{on } Q_T, \qquad b(t) = a(T-t), \tag{39}$$

$$z(0,.) = 0, \qquad z_{|\partial\Omega} = 0. \tag{40}$$

It turns out that (38) is valid for all $q \geq 2$. Indeed, for $q = \infty$, it is an easy consequence of the maximum principle which holds for (39). For $q = 2$, we may (formally) multiply (39) by $-\Delta z$ to obtain

$$\frac{\partial}{\partial t}\frac{1}{2}\int_\Omega |\nabla z(t)|^2 + \int_\Omega b(\Delta z)^2 = -\int_\Omega \theta\Delta z.$$

Using the uniform estimate (36) and Young's inequality, we deduce that

$$\frac{1}{2}\int_\Omega |\nabla z(t)|^2 + \min(d_1, d_2)\int_0^t\int_\Omega (\Delta z)^2 \leq \frac{1}{2}\min(d_1, d_2)\int_0^t\int_\Omega (\Delta z)^2 + C\int_0^t\int_\Omega \theta^2,$$

which implies

$$\int_{Q_T} ||\Delta z||_{L^2(Q_T)} \leq C||\theta||_{L^2(Q_T)}, \tag{41}$$

where C depends only on d_1 and d_2. Therefore, estimate (38) is true for $q = 2$ (and, by interpolation, for any $q \in [2, \infty]$). We even have that z_t and Δz (and not only z itself) are bounded in $L^2(Q_T)$.

As a consequence of the counterexamples in Theorems 2.1 and 2.2, we will see that (38) is false when q is close to 1. More precisely, we have the following.

Proposition 2.3. *There exists $p \in (1, 2)$, $b_n \in C^\infty(\overline{Q}_T)$, $d_1, d_2 > 0$, $\theta_n \in C^\infty(\overline{Q}_T)$, $\theta_n \geq 0$, and $z_n \in C^\infty(\overline{Q}_T)$, the solution of*

$$\frac{\partial z_n}{\partial t} - b_n\Delta z_n = \theta_n \quad \textit{on } Q_T, \tag{42}$$

$$z_n(0,.) = 0, \qquad z_{n_{|\partial B}} = 0 \tag{43}$$

such that

$$0 < d_1 \leq b_n \leq d_2, \tag{44}$$

$$||\theta_n||_{L^p(Q_T)} = 1, \tag{45}$$

$$\lim_{n\to\infty} ||z_n||_{L^\infty(0,T;L^1(B))} = +\infty. \tag{46}$$

Finally, we mention a final consequence of independent interest for the following nonlinear Hamilton–Jacobi evolution equations.

Proposition 2.4. *There exist $p \in (1, 2)$, $d_1, d_2 > 0$, $\theta_n \in C^\infty(\overline{Q}_T)$, and $z_n \in L^2(Q_T)$ with $\Delta_{z_n}, \partial z_n/\partial t \in L^2(Q_T)$ the solution of*

$$\frac{\partial z_n}{\partial t} - \max(d_1\Delta z_n, d_2\Delta z_n) = \theta_n \quad \textit{on } Q_T, \tag{47}$$

$$z_n(0,.) = 0, \qquad z_{n_{|\partial B}} = 0 \tag{48}$$

such that

$$||\theta_n||_{L^p(Q_T)} = 1, \tag{49}$$

$$\lim_{n\to\infty} ||z_n||_{L^\infty(0,T;L^1(B))} = +\infty. \tag{50}$$

3. The proofs.

Preliminary remarks. As stated in section 2, the proofs of Theorems 2.1 and 2.2 are obtained by constructing explicit functions u and v satisfying the inequality

$$u_t - d_1\Delta u + v_t - d_2\Delta v \leq 0. \tag{51}$$

It will turn out that they are also solutions of systems of the form (16), (17) and (24), (25) with the properties listed in the statements of the theorems.

For reasons that come from a precise analysis of the problem and a guess of the possible singularities, we look a priori for functions u and v of the form

$$u(t,x) = \frac{a(T-t)+b|x|^2}{(T-t+|x|^2)^\gamma}, \qquad v(t,x) = \frac{c(T-t)+d|x|^2}{(T-t+|x|^2)^\gamma}, \tag{52}$$

where $|.|$ denotes the Euclidian norm and $a, b, c, d > 0$ and $\gamma > 1$ are to be determined so that (51) holds for some $d_1, d_2 > 0$, also to be determined. This has been done with the help of the formal computation software Maple, where the unknown coefficients can be progressively adapted “by hand” to satisfy (51). As a consequence, the solutions that we found in this way are not numerically simple.

Here we give one solution which can also be explicitly checked by direct computations. For this, we choose

$$N = 10 \text{ (for the dimension)}, \qquad \gamma = 5/4, \tag{53}$$

$$a = 1/25, \qquad b = 1, \qquad c = 11/2, \qquad d = 1/10, \tag{54}$$

$$d_1 = 1, \qquad d_2 = 1/10. \tag{55}$$

Lengthy but straightforward computations show that u and v given by (52) with the data (53)–(55) satisfy inequality (51) and, more precisely,

$$u_t - d_1\Delta u = \frac{A_1(T-t)^2 + B_1(T-t)|x|^2 + C_1|x|^4}{(T-t+|x|^2)^{\gamma+2}}, \tag{56}$$

$$v_t - d_2\Delta v = \frac{A_2(T-t)^2 + B_2(T-t)|x|^2 + C_2|x|^4}{(T-t+|x|^2)^{\gamma+2}}, \tag{57}$$

$$u_t + v_t - d_1\Delta u - d_2\Delta v = \frac{A(T-t)^2 + B(T-t)|x|^2 + C|x|^4}{(T-t+|x|^2)^{\gamma+2}}, \tag{58}$$

where

$$A_1 = -1899/100, \qquad B_1 = -323/100, \qquad C_1 = 496/100, \tag{59}$$

$$A_2 = 1194/80, \qquad B_2 = 281/80, \qquad C_2 = -427/80, \tag{60}$$

$$A = -1626/400, \qquad B = 113/400, \qquad C = -151/400. \tag{61}$$

We check that $B^2 - 4AC < 0$ so that (51) holds.

Remark. Note that other explicit solutions are also found in dimension $N = 1$. We easily show that they satisfy (24) and (25) in Theorem 2.2. However, the corresponding functions f and g of Theorem 2.1 are technically difficult to present.

TABLE 1

k	$\tilde{\lambda_k}$	$\tilde{\mu_k}$	$\tilde{\nu_k}$
0	$2^{-3} \times 3^7 \times 5^8 \times 43 \times 653$	$-2^{-2} \times 3^7 \times 5^9 \times 3023$	$-2^{-3} \times 3^9 \times 5^8 \times 239$
1	$2^{-2} \times 3^5 \times 5^8 \times 7 \times 7703$	$-2^{-4} \times 3^5 \times 5^8 \times 11 \times 20717$	$-2^{-4} \times 3^5 \times 5^8 \times 12203$
2	$-2 \times 3^5 \times 5^7 \times 19 \times 359$	$2^{-3} \times 3^4 \times 5^7 \times 29 \times 61 \times 131$	$-2^{-3} \times 3^4 \times 5^7 \times 7 \times 79 \times 173$
3	$-2^4 \times 3 \times 5^6 \times 23 \times 31 \times 397$	$3 \times 5^6 \times 409 \times 9011$	$-3^2 \times 5^6 \times 281159$
4	$-3^{-1} \times 2^7 \times 5^5 \times 61 \times 4567$	$3^{-1} \times 2^3 \times 5^8 \times 28591$	$-3^{-1} \times 2^3 \times 5^5 \times 883517$
5	$-3^{-1} \times 2^{13} \times 5^3 \times 17 \times 19^2$	$3^{-1} \times 2^6 \times 5^4 \times 59 \times 2087$	$-3^{-1} \times 2^6 \times 5^3 \times 13 \times 73 \times 179$

Proof of Theorem 2.1. We will show that u and v defined as above satisfy the conclusions of the theorem. Obviously, $u(0), v(0) \in C^\infty(\bar{B})$, $u_{|\partial B} = \alpha_1 \in C^\infty([0,T])$, $v_{|\partial B} = \alpha_2 \in C^\infty([0,T])$, and, since $\gamma > 1$,

$$\lim_{t \uparrow T} ||u(t)||_\infty = \lim_{t \uparrow T} ||v(t)||_\infty = +\infty. \tag{62}$$

More precisely, all of the L^p-norms for $p \geq 20$ blow up at $t = T$.

Now we need to determine f and g. Starting from (56) and (57), we look for polynomial functions P and Q, homogeneous and of degree 5 in u and v, such that

$$u_t - d_1 \Delta u = P(u,v), \qquad v_t - d_2 \Delta v = Q(u,v). \tag{63}$$

Formal computations lead to the following expressions, which can be checked directly by, again, lengthy but straightforward computations:

$$P(u,v) = \sum_{k=0}^{5} \lambda_k u^{5-k} v^k, \qquad Q(u,v) = \sum_{k=0}^{5} \mu_k u^{5-k} v^k \tag{64}$$

with $\lambda_k = (229)^{-5} \tilde{\lambda}_k$ and $\mu_k = (229)^{-5} \tilde{\mu}_k$ and where $\tilde{\lambda}_k$ and $\tilde{\mu}_k$ are given in Table 1.

We check that $P + Q = \sum_{k=0}^{5} \nu_k u^{5-k} v^k$, where all the $\nu_k = (229)^{-5} \tilde{\nu}_k$'s are negative so that for λ close to 1, $P + \lambda Q$ also has negative coefficients. Therefore, (20) holds for P and Q. However, P and Q do not satisfy (22). Since u and v are bounded from below on Q_T by

$$m_1 = \frac{\min(a,b)}{(T+1)^{\gamma-1}}, \qquad m_2 = \frac{\min(c,d)}{(T+1)^{\gamma-1}},$$

respectively, we can always modify P and Q on a neighborhood of $\{0\} \times [0, \infty[\, \cup \, [0, \infty[\, \times \{0\}$. Let $\varphi \in C^\infty(\mathbb{R}^2)$ be a function which is identically one on $[m_1, \infty) \times [m_2, \infty)$ and such that

$$\forall u, v \geq 0, \quad \varphi(0, v) = \varphi(u, 0) = 0.$$

We set

$$\forall u, v \geq 0, \quad f(u,v) := \varphi(u,v)P(u,v), \qquad g(u,v) := \varphi(u,v)Q(u,v).$$

Then for the same values of λ close to 1 as above,

$$f + \lambda g = \varphi(P + \lambda Q) \leq 0,$$

so (16), (17), (20), (21), and (22) are satisfied.

Proof of Theorem 2.2. We use the same functions u and v as above. We only have to prove that the expressions in (56) and (57) can be written as in (24) and (25). For this we choose $\alpha, \beta > 1$ such that

$$\alpha + \beta > (\gamma + 2)/(\gamma - 1) \tag{65}$$

and (see (52), (56), and (57))

$$c_1(t,x) := \frac{A_1(T-t)^2 + B_1(T-t)|x|^2 + C_1|x|^4}{(a(T-t) + b|x|^2)^\alpha (c(T-t) + d|x|^2)^\beta} (T - t + |x|^2)^{(\alpha+\beta)\gamma - \gamma - 2},$$

$$c_2(t,x) := \frac{A_2(T-t)^2 + B_2(T-t)|x|^2 + C_2|x|^4}{(a(T-t) + b|x|^2)^\alpha (c(T-t) + d|x|^2)^\beta} (T - t + |x|^2)^{(\alpha+\beta)\gamma - \gamma - 2}.$$

Obviously, with this choice of c_1 and c_2, relations (24) and (25) hold. The sign of $c_1 + c_2$ is the same as the sign of $A(T-t)^2 + B(T-t)|x|^2 + C|x|^4$ in (58), which has already been checked to be negative. Now, thanks to the choice of (6.5), c_1 and c_2 are at least continuous on $\overline{Q_T}$. Actually, they are C^∞ except at the point $(T, 0)$, where they tend to 0. We can make them more regular by choosing α and β even larger.

Proof of Proposition 2.3. We take u and v as in (52) and we set

$$w_n(t,x) := u\left(t - \frac{1}{n}, x\right) + v\left(t - \frac{1}{n}, x\right), \tag{66}$$

$$a_n(t,x) := \left[d_1 u\left(t - \frac{1}{n}, x\right) + d_2 v\left(t - \frac{1}{n}, x\right)\right] / w_n(t,x). \tag{67}$$

Note that

$$0 < \min(d_1, d_2) \leq a_n \leq \max(d_1, d_2).$$

By (51),

$$\frac{\partial w_n}{\partial t} - \Delta(a_n w_n) \leq 0 \quad \text{ on } Q_T. \tag{68}$$

For $\theta \in C^\infty(\overline{Q}_T)$, $\theta \geq 0$, and $b_n(t) = a_n(T-t)$, let $z \in C^\infty(\overline{Q}_T)$, $z \geq 0$ be the solution of

$$\frac{\partial z}{\partial t} - b_n \Delta z = \theta \quad \text{ on } Q_T, \tag{69}$$

$$z(0,.) = 0, \qquad z_{|\partial B} = 0. \tag{70}$$

The existence of z is classical (see, e.g., [5]). We set $\tilde{z}(t) = z(T-t)$ so that

$$-(\tilde{z}_t + a_n \Delta \tilde{z}) = \tilde{\theta}, \qquad \tilde{\theta}(t) = \theta(T-t). \tag{71}$$

Now let $\varphi \in C_0^\infty(B)$ with $0 \le \varphi \le 1$ and $\varphi \equiv 1$ on $B(0,1/2)$. Using (68)–(71), we have

$$\begin{aligned}\int_{Q_T} \tilde{\theta} w_n \varphi = \int_{Q_T} -w_n \varphi(\tilde{z}_t + a_n \Delta \tilde{z}) = \int_\Omega w_n(0)\varphi \tilde{z}(0) + \int_{Q_T} \tilde{z}(w_{n_t}\varphi - \Delta(a_n w_n \varphi))\\ = \int_\Omega w_n(0)\varphi z(T) + \int_{Q_T} \tilde{z}[\varphi(w_{n_t} - \Delta(a_n w_n)) - 2\nabla\varphi\nabla(a_n w_n) - a_n w_n \Delta\varphi].\end{aligned}$$

Since $\nabla\varphi$ and $\Delta\varphi$ are identically zero around the origin, the terms $\nabla\varphi\nabla(a_n w_n)$ and $a_n w_n \Delta\varphi$ are uniformly bounded independently of n. Also using inequality (68), we obtain

$$\begin{aligned}\int_{Q_T} \tilde{\theta} w_n \varphi &\le \int_\Omega w_n(0) z(T) + C\int_{Q_T} z\\ &\le (\|w_n(0)\|_\infty + CT)\|z\|_{L^\infty(0,T,L^1(B))}.\end{aligned} \tag{72}$$

If we had

$$\|z\|_{L^\infty(0,T,L^1(B))} \le k\|\theta\|_{L^p(Q_T)} \tag{73}$$

for some $p \in (1,2)$, some $k = k(d_1, d_2, p, T)$, and all $\theta \in C^\infty(\overline{Q}_T)$, then from (72) we would deduce by duality that for some C independent of n,

$$\|w_n\varphi\|_{L^{p'}(Q_T)} \le C.$$

This is false for p' large enough (that is, for p small enough) by the construction of u and v and the definition (66) of w_n. Therefore, the solutions of (69) and (70) do not satisfy estimate (73) (see (72)) for some p close to 1, whence the statement of Proposition 2.3 follows.

Proof of Proposition 2.4. For $\theta_n \in C^\infty(\overline{Q}_T)$, it is classical (see, e.g., [1] and [10]) that there exists a solution of (47) and (48) whose regularity is at least such that

$$z_n \in L^\infty(Q_T), \qquad z_{n_t}, \Delta z_n \in L^2(Q_T),$$

and (47) is satisfied at least a.e. $x, t \in Q_T$. We choose θ_n as in the statement of Proposition (2.3). We obviously have that

$$\max(d_1\Delta z_n, d_2\Delta z_n) \ge b_n \Delta z_n \quad \text{a.e.}$$

so that

$$\frac{\partial z_n}{\partial t} - b_n \Delta z_n \ge \theta_n.$$

By the maximum principle applied to the operator $\partial/\partial t - b_n\Delta$, this solution z_n is greater than the one defined in Proposition 2.3. As a consequence, (50) follows from (46).

REFERENCES

[1] M. Drissi-Zelladji, *Méthodes d'agrégation et méthodes des familles*, thèse de doctorat, Université de Franche-Comté, Besancon, France, 1994.

[2] S. Hollis, R. H. Martin, and M. Pierre, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744–761.

[3] S. Hollis and J. Morgan, *Interior estimate for a class of reaction-diffusion systems from L^1 a priori estimate*, J. Differential Equations, 98 (1992), pp. 260–276.

[4] A. Haraux and A. Youkana, *On a result of K. Masuda concerning reaction-diffusion equations*, Tôhoku Math. J., 40 (1988), pp. 159–163.

[5] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs 23, AMS, Providence, 1968.

[6] R. H. Martin and M. Pierre, *Nonlinear reaction diffusion systems*, in Nonlinear Equations in the Applied Sciences, W. F. Ames and C. Rogers, eds., Mathematics in Science and Engineering 185, Academic Press, New York, 1991.

[7] K. Masuda, *On the global existence and asymptotic behavior of reaction-diffusion equations*, Hokkaido Math. J., 12 (1983), pp. 360–370.

[8] J. Morgan, *Global existence for semilinear parabolic systems*, SIAM J. Math. Anal., 20 (1989), pp. 1128–1144.

[9] M. Pierre and D. Schmitt, *Global existence for a reaction-diffusion system with a balance law*, in Semigroups of Linear and Nonlinear Operators and Applications (Curaçao, 1992), G. R. Goldstein and G. A. Goldstein, eds., Kluwer Academic Publishers, Norwell, MA, 1993.

[10] D. Schmitt, *Existence globale ou explosion pour les systèmes de réaction-diffusion avec contrôle de masse*, thèse d'université, Université Henri Poincaré–Nancy I, Vandoeuvre-lès-Nancy, France, 1995.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 270–273, March 1997

HIGHER GRADIENT INTEGRABILITY OF EQUILIBRIA FOR CERTAIN RANK-ONE CONVEX INTEGRALS*

MICHAEL M. DOUGHERTY[†] AND DANIEL PHILLIPS[‡]

Abstract. In this note, we prove $L^m_{\text{loc}}(\Omega)$, $1 < m < \infty$, estimates of the gradients of equilibria $\mathcal{U} \in W^{1,p}(\Omega; \mathbb{R}^N)$ for a class of variational integrals $\int_\Omega f(D\mathcal{U})\,dx$. Our proof employs a bootstrap argument based on a result of DiBenedetto and Manfredi [*Amer. J. Math.*, 115 (1993), pp. 1107–1134].

Key words. rank-one convex, calculus of variations

AMS subject classifications. Primary, 49N60; Secondary, 35J60, 73C50

PII. S0036141095293384

1. Introduction and main result.

Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$, be a bounded domain. For $\mathcal{U} \in W^{1,p}(\Omega; \mathbb{R}^N)$, $1 < p < \infty$, define

$$\mathcal{F}(\mathcal{U}) = \int_\Omega f(D\mathcal{U})dx, \tag{1.1}$$

where f has the form

$$f(A) = |A|^p + g(A) \tag{1.2}$$

and g is a rank-one convex function such that

$$\bigtriangledown g(A) \text{ exists} \quad \text{for all } A \in \mathcal{M}^{N\times n}, \tag{1.3}$$

$$|g(A)| \leq a|A|^q + b \tag{1.4}$$

for real numbers a, b, and $1 \leq q < p < \infty$. By the rank-one convexity of g, we mean that

$$g(\lambda A + (1-\lambda)B) \leq \lambda g(A) + (1-\lambda)g(B) \tag{1.5}$$

holds for all $0 \leq \lambda \leq 1$ and $A, B \in \mathcal{M}^{N\times n}$ with rank $\{A - B\} \leq 1$. Our result concerns equilibrium points of $\mathcal{F}$, i.e., functions $\mathcal{U} \in W^{1,p}(\Omega; \mathbb{R}^N)$ which satisfy

$$\text{div}\left(\frac{\partial f}{\partial A}(D\mathcal{U})\right) = 0 \quad \text{in } \mathcal{D}'(\Omega; \mathbb{R}^N). \tag{1.6}$$

We prove the following.

THEOREM. *Suppose $\mathcal{U} \in W^{1,p}(\Omega; \mathbb{R}^N)$, $1 < p < \infty$, is an equilibrium point of* (1.1) *and that* (1.2)–(1.4) *hold. Then for each subset $\Omega' \Subset \Omega$ and every $m \in (1, \infty)$, there exists a constant $C = C(a, b, m, p, n, N, \Omega', \Omega)$ such that*

$$\|D\mathcal{U}\|_{L^m(\Omega')} \leq C\big(1 + \|D\mathcal{U}\|_{L^p(\Omega)} + \|\mathcal{U}\|_{L^p(\Omega)}\big). \tag{1.7}$$

* Received by the editors October 23, 1995; accepted for publication December 4, 1995.
http://www.siam.org/journals/sima/28-2/29338.html

[†] Department of Mathematics, Pennsylvania State University, Berks Campus, Reading, PA 19610 (mmd5@psu.edu).

[‡] Department of Mathematics, Purdue University, West Lafayette, IN 47907 (phillips@euclid.math.purdue.edu). The research of this author was partially supported by NSF grant DMS-9306199.

Remark 1. In particular, the result holds for minimizers of (1.1)–(1.5) when they exist, i.e., functions $\mathcal{U} \in W^{1,p}(\Omega; \mathbb{R}^N)$ such that $\mathcal{F}(\mathcal{U}) \leq \mathcal{F}(\mathcal{V})$ whenever $\mathcal{U} - \mathcal{V} \in W_0^{1,p}(\Omega; \mathbb{R}^N)$.

Remark 2. By the Sobolev embedding theorem, we have $\mathcal{U} \in C^\alpha(\Omega; \mathbb{R}^N)$ for all $\alpha \in (0,1)$. Previously, local Hölder continuity of minimizers for $0 < \alpha < 1$ was established in [F].

2. Background and examples. Rank-one convex functions satisfying (1.2)–(1.4) arise in optimization problems [K-S] and as approximate energies for models from nonlinear elasticity [B-P]. Such functions $f(A)$ have the feature that they can depend nonlinearly on the subdeterminants of A. For example, if we let $n = N$ and $g(A) = h(\det A)$ with h convex, then g is rank-one convex. If $p > n$ and

$$|h(d)| \leq C(|d| + 1), \tag{2.1}$$

then

$$f(A) = |A|^p + h(\det A) \tag{2.2}$$

is an example of an integrand satisfying (1.2)–(1.4). Functionals of this type but with a much more singular dependence on d are used in nonlinear elasticity theory. There one assumes $h(d) = +\infty$ for $d \leq 0$. Existence of minimizers for the singular problem is established but little is known about the minimizers themselves. One approach in studying the elasticity problem is to consider a sequence of approximate problems with

$$|h_k(d)| \leq c_k(|d| + 1) \tag{2.3}$$

and

$$h_k(d) \uparrow h(d) \quad \text{as } k \longrightarrow \infty. \tag{2.4}$$

These problems do have equilibrium points. The motivation for the present work is the investigation of gradient estimates for solutions to the approximate problem.

Prior work on estimating the gradient of solutions to such problems was done in [C-E] and [G-M]. In these papers, it was assumed that f is C^2 and

$$\lim_{|A| \to \infty} \frac{D^2(|A|^p) - D^2 f(A)}{|A|^{p-2}} = 0, \tag{2.5}$$

where $p \geq 2$. These works show that minimizers have locally bounded gradients. However, for a large class of integrands (such as (2.1)–(2.2), for instance), one must have $p > 2n$ in order for (2.5) to hold. To see this for our example, let $d \in \mathbb{R}$ such that $h''(d) > 0$ and $\lambda > 0$ and set

$$A = \operatorname{diag}[A_{11}, \ldots, A_{nn}], \tag{2.6}$$

where $A_{11} = d\lambda^{1-n}$ and $A_{ii} = \lambda$ for $2 \leq i \leq n$. Then

$$\left. \frac{\partial^2 h(\det A)}{\partial A_{11} \partial A_{11}} \right|_A = h''(d)\lambda^{2n-2}. \tag{2.7}$$

If we then consider (2.5) with $A = A(\lambda)$ and let $\lambda \longrightarrow \infty$, it follows that $p > 2n$ is required.

Thus for many examples with $n < p \leq 2n$, our theorem provides L^m estimates for the gradient of solutions for all $m < \infty$. Moreover, this information is new even for minimizers. In cases where (1.2)–(1.4) hold, our result establishes L^m estimates for equilibria in general.

3. Lemmas and proof of theorem. The proof will utilize the following lemmas.

Lemma 3.1. *Given* (1.3)–(1.5)*, there exists a constant $C < \infty$ such that*

$$|\bigtriangledown g(A)| \le C(1 + |A|^{q-1}) \tag{3.1}$$

holds for all $A \in \mathcal{M}^{N\times n}$.

Lemma 3.2. *If $\mathcal{U} \in W^{1,p}(\Omega;\mathbb{R}^N)$, $1 < p < \infty$, is an equilibrium point of* (1.1)–(1.5)*, then $\mathcal{U}$ satisfies the Euler–Lagrange equations in the weak sense:*

$$p \cdot \operatorname{div}(|D\mathcal{U}|^{p-2}D\mathcal{U}) = -\operatorname{div}(\bigtriangledown g(D\mathcal{U})). \tag{3.2}$$

That is, for each $\varphi \in W_0^{1,p}(\Omega;\mathbb{R}^N)$,

$$\int_\Omega p|D\mathcal{U}|^{p-2}\left\langle D\mathcal{U}, D\varphi\right\rangle dx = -\int_\Omega \left\langle \bigtriangledown g(D\mathcal{U}), D\varphi\right\rangle dx. \tag{3.3}$$

The first lemma can be found in [Dac, section 4.2.1.2]. The second lemma is a straightforward application of the dominated convergence theorem, (3.1), and Hölder's inequality. Our final lemma is a result of DiBenedetto and Manfredi [DB-M, Theorem 2.1] (see also Iwaniec [I, Theorem 2]).

Lemma 3.3. *Let $F \in L^p(\Omega;\mathbb{R}^{nN})$. Suppose $\mathcal{U} \in W^{1,p}(\Omega;\mathbb{R}^N)$ is a weak solution of*

$$\operatorname{div}(|D\mathcal{U}|^{p-2}D\mathcal{U}) = \operatorname{div}(|F|^{p-2}F) \quad \text{in } \Omega, \tag{3.4}$$

i.e.,

$$\int_\Omega |D\mathcal{U}|^{p-2}\langle D\mathcal{U}, D\varphi\rangle dx = \int_\Omega |F|^{p-2}\langle F, D\varphi\rangle dx \tag{3.5}$$

for all $\varphi \in W_0^{1,p}(\Omega;\mathbb{R}^N)$. Assume $d > 0$ and B_d, B_{2d}, $B_{4d} \subset \Omega$ are concentric balls of radii d, $2d$, and $4d$, respectively. Let $\eta \in C_0^\infty(B_{2d})$, $0 \le \eta \le 1$, and $\eta \equiv 1$ in B_d. If $|F| \in L^r_{\mathrm{loc}}(\Omega)$ for some $r \in [p,\infty)$, then $|D\mathcal{U}| \in L^r_{\mathrm{loc}}(\Omega)$, and for each $B_{4d} \subset \Omega$, there exists $\gamma = \gamma(n,N,p,r,d)$ such that

$$\|D(\eta\mathcal{U})\|_{L^r(\Omega)} \le \gamma\big(\|\eta F\|_{L^r(\Omega)} + \|\eta\mathcal{U}\|_{L^p(\Omega)}\big). \tag{3.6}$$

Proof of Theorem. We begin by rewriting the Euler–Lagrange equations (3.2) in the form of (3.4). We do this by taking

$$|F|^{p-2}F = -\frac{1}{p}\bigtriangledown g(D\mathcal{U}),$$

that is,

$$F = -\left(\frac{1}{p}\right)^{\frac{1}{p-1}} |\bigtriangledown g(D\mathcal{U})|^{\frac{2-p}{p-1}} \bigtriangledown g(D\mathcal{U}). \tag{3.7}$$

In particular,

$$|F| = \left(\frac{1}{p}\right)^{\frac{1}{p-1}} |\bigtriangledown g(D\mathcal{U})|^{\frac{1}{p-1}}. \tag{3.8}$$

Without loss of generality, assume $q > 1$. From (3.1), we have that

$$|\bigtriangledown g(D\mathcal{U})| \leq C(1 + |D\mathcal{U}|^{q-1}) \in L^{\frac{p}{q-1}}(\Omega). \tag{3.9}$$

Thus

$$|F| \in L^{(p-1)\cdot\frac{p}{q-1}}(\Omega). \tag{3.10}$$

By Lemma 3.3, we have

$$|D\mathcal{U}| \in L_{\text{loc}}^{p\cdot\frac{p-1}{q-1}}(\Omega), \tag{3.11}$$

and an estimate from (3.6). Note that $r \equiv p \cdot \frac{p-1}{q-1} > p$.

The other gradient estimates follow from the first by a bootstrap argument. Since we have proved that $|D\mathcal{U}| \in L_{\text{loc}}^{p\cdot(p-1)/(q-1)}(\Omega)$, we can further say that

$$|\bigtriangledown g(D\mathcal{U})| \in L_{\text{loc}}^{p\cdot\frac{p-1}{q-1}\cdot\frac{1}{q-1}}(\Omega), \tag{3.12}$$

which implies that

$$|F| \in L_{\text{loc}}^{p\cdot(\frac{p-1}{q-1})^2}(\Omega), \tag{3.13}$$

which by Lemma 3.3 gives

$$|D\mathcal{U}| \in L_{\text{loc}}^{p\cdot(\frac{p-1}{q-1})^2}(\Omega). \tag{3.14}$$

By induction, we get $|D\mathcal{U}| \in L_{\text{loc}}^{m}(\Omega)$ for each $m \in (p, \infty)$. □

REFERENCES

[B-P] P. Bauman and D. Phillips, *Univalent minimizers of polyconvex functionals in two dimensions*, Arch. Rational Mech. Anal., 126 (1994), pp. 161–181.

[C-E] M. Chipot and L. Evans, *Linearisation at infinity and Lipschitz estimates for certain problems in the calculus of variations*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 291–303.

[Dac] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[DB-M] E. DiBenedetto and J. Manfredi, *On the higher integrability of the gradient of weak solutions of certain degenerate elliptic systems*, Amer. J. Math., 115 (1993), pp. 1107–1134.

[F] M. Fuchs, *Regularity for a class of variational integrals motivated by non-linear elasticity*, Asymptotic Anal., 9 (1994), pp. 23–38.

[G-M] M. Giaquinta and G. Modica, *Remarks on the regularity of the minimizers of certain degenerate functionals*, Manuscripta Math, 57 (1986), pp. 55–99.

[I] T. Iwaniec, *Projections onto gradient fields and L^p-estimates for degenerate elliptic equations*, Studia Math., 75 (1983), pp. 293–312.

[K-S] R. Kohn and G. Strang, *Explicit relaxation of a variational problem in optimal design*, Bull. Amer. Math. Soc., 9 (1983), pp. 211–214.

SIAM J. Math. Anal.
Vol. 28, No. 2, pp. 274–303, March 1997

THE LIMIT OF THE FULLY ANISOTROPIC DOUBLE-OBSTACLE ALLEN–CAHN EQUATION IN THE NONSMOOTH CASE*

CHARLES M. ELLIOTT[†] AND REINER SCHÄTZLE[‡]

Abstract. In this paper, we prove that solutions of the anisotropic Allen–Cahn equation in double-obstacle form with kinetic term

$$\varepsilon\beta(\nabla\varphi)\partial_t\varphi - \varepsilon\nabla A'(\nabla\varphi) - \frac{1}{\varepsilon}\varphi = \frac{\pi}{4}u \quad \text{in } [|\varphi| < 1],$$

where A is a convex function homogeneous of degree two and β depends only on the direction of $\nabla\varphi$, converge to an anisotropic mean-curvature flow

$$\beta(N)V_N = -\mathrm{tr}(B(N)D^2B(N)R) - B(N)u.$$

Here N, V_N, and R denote the normal, the normal velocity, and the second fundamental form of the interface, respectively, and $B := \sqrt{2A}$.

Key words. anisotropic Allen–Cahn equation, mean-curvature flow, viscosity solutions, double-obstacle problem

AMS subject classifications. 35R35, 35K22

PII. S0036141095286733

1. Introduction.

The Allen–Cahn equation

$$\varepsilon\partial_t\varphi_\varepsilon - \varepsilon\Delta\varphi_\varepsilon + \frac{1}{\varepsilon}W'(\varphi_\varepsilon) = c_W u, \tag{1}$$

where $W(t) := (t^2-1)^2$ is a double well potential and c_W is a certain constant depending only on W, was introduced by Allen and Cahn in [1] as a model for grain boundary motion. It is the L^2-gradient flow of the energy functional

$$F_\varepsilon(\varphi) := \int \frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{\varepsilon}W(\varphi) - c_W u\varphi. \tag{2}$$

In its double-obstacle form (see Blowey and Elliott in [5]), the Allen–Cahn equation is replaced by a parabolic variational inequality with constraint that reads

$$\begin{gathered}\forall\eta \in \mathrm{L}^2(\mathrm{H}^{1,2}),\ |\eta| \le 1 : \\ \int \varepsilon\partial_t\varphi\ (\varphi-\eta) + \varepsilon\nabla\varphi\ (\nabla\varphi - \nabla\eta) - \frac{1}{\varepsilon}\varphi(\varphi-\eta) - \frac{\pi}{4}u(\varphi-\eta) \le 0, \\ |\varphi| \le 1.\end{gathered} \tag{3}$$

The convergence of solutions of the Allen–Cahn equation to the mean-curvature flow

$$V_N = -\kappa - u \tag{4}$$

was proved by Chen in [6], de Mottoni and Schatzmann in [10], Evans, Soner, and Souganidis in [13], and Bellettini and Paolini in [3]; in double-obstacle form, this was

* Received by the editors May 26, 1995; accepted for publication (in revised form) January 9, 1996.
http://www.siam.org/journals/sima/28-2/28673.html

[†] Centre for Mathematical Analysis and Its Applications, School of Mathematical Sciences, University of Sussex, Falmer, Brighton BN1 9QH, United Kingdom (c.m.elliott@sussex.ac.uk).

[‡] Institut für Angewandte Mathematik, Universität Bonn, Wegelerstraße 6, D-53115 Bonn, Germany (reiner@iam.uni-bonn.de). The research of this author was supported by the ESF.

proved by Chen and Elliott in [8], Nochetto, Paolini, and Verdi in [17] and Nochetto and Verdi in [18].

Anisotropy is now introduced by replacing the gradient in (2) and considering the functional

$$F_\varepsilon(\varphi) := \int \varepsilon A(\nabla\varphi) + \frac{1}{\varepsilon}W(\varphi) - c_W u\varphi, \tag{5}$$

where A is convex and homogeneous of degree two. Moreover, we do not consider the gradient flow of this functional but introduce a kinetic factor depending on $\nabla\varphi$ in front of $\partial_t\varphi$. Then the Allen–Cahn equation takes the form

$$\varepsilon\beta(\nabla\varphi)\partial_t\varphi_\varepsilon - \varepsilon\nabla A'(\nabla\varphi_\varepsilon) + \frac{1}{\varepsilon}W'(\varphi_\varepsilon) = c_W u, \tag{6}$$

where β is a positive function homogeneous of degree zero. McFadden et al. [15], Wheeler and McFadden [21], and Bellettini and Paolini [4] with $\beta = 1$ used formal asymptotics to provide the conjecture that (6) approximates the anisotropic mean-curvature flow, which reads, in two space dimensions,

$$\beta(N)V_N = -\gamma(\gamma + \gamma'')(\theta(N))\kappa - \gamma(\theta(N))u, \tag{7}$$

where γ is 2π-periodic, θ is the angle, and $\gamma(\theta(N)) := \sqrt{2A(N)}$. This conjecture was proved by the authors in [11] without the kinetic factor in the case when the evolution of the anisotropic mean-curvature flow is smooth. Moreover, it is proved there that the Hausdorff distance between the zero-level set of φ_ε, the solution of (6) in its double-obstacle form, and the interface of the flow is of order $O(\varepsilon^2)$. In the isotropic case, this bound was established by Nochetto, Paolini, and Verdi in [17].

The first difficulty that arises from (6) is how to define a weak solution for this equation. The problem is that, unless β is constant, β is discontinuous at 0. Thus far, it is not clear how to give such a definition. Instead, we consider (6) and its double-obstacle variant in the viscosity sense. For viscosity solutions see, for example, the article [9] of Crandall, Ishi, and Lions. In spite of (6) not admitting a comparison principle in the presence of a nonconstant kinetic factor, we will prove that it has a solution. This will be done by proving uniform convergence of solutions of regularized equations to (6) in section 2.

In section 4, we prove that solutions of (6) in double-obstacle form approximate the corresponding anisotropic mean-curvature flow. The existence of a level-set solution describing this flow was proved by Chen, Giga, and Goto in [7]. The approximation is then meant in the sense that φ_ε, the solution of (6) in double-obstacle form, converges to 1 (respectively, to -1) where the level-set solution is positive (respectively, negative). This is proved by constructing suitable sub- and supersolutions of (6) in double-obstacle form. The construction is similar to that given by Nochetto and Verdi in [18]. Because of the anisotropic nature of the problem, instead of the ordinary distance function, we had to use a distance function that is induced by a Finsler geometry as outlined by Bellettini and Paolini in [4]; see section 3. These sub- and supersolutions are finally compared with the solutions, yielding our result.

The convergence result determines the limit of φ_ε uniquely when the interface of the flow does not develop an interior. Otherwise—that is, in the case called *fattening*—there remains an ambiguity.

For applications, it may be easier to consider equations where the kinetic term does not have a discontinuity at 0. After completing this paper, in [12], we considered

$\beta_\varepsilon \in \mathrm{C}^0(\mathbb{R}^n)$ with $\beta_\varepsilon(p) = \beta(p)$ for $|p| \geq \varepsilon$. The corresponding Allen–Cahn equation reads

$$\varepsilon \beta_\varepsilon(\tilde{\varphi}_\varepsilon)\partial_t \tilde{\varphi}_\varepsilon - \varepsilon \nabla A'(\nabla \tilde{\varphi}_\varepsilon) + \frac{1}{\varepsilon} W(\tilde{\varphi}_\varepsilon) = c_W u.$$

In [12], we proved that the Hausdorff distance between the zero-level set of the solutions of its double-obstacle variant and the interface of the flow is of order $O(\varepsilon^2)$ when this flow is smooth.

2. Existence and comparison.

2.1. Notation. Let $\beta \in C^{0,1}_{\mathrm{loc}}(\mathbb{R}^n - \{0\})$, $A \in C^{2,1}_{\mathrm{loc}}(\mathbb{R}^n - \{0\})$, and $u \in W^{2,1}_\infty(\mathbb{R}^n \times [0,T])$ be given. We assume that β is homogeneous of degree zero and that A is homogeneous of degree two. Set $B := \sqrt{2A}$. We assume the following bounds:

$$\begin{gathered} n,\ T,\ \|\beta\|_{C^{0,1}(B_2(0)-B_{1/2}(0))},\ \|B\|_{C^{2,1}(B_2(0)-B_{1/2}(0))}, \\ \|A\|_{C^{2,1}(B_2(0)-B_{1/2}(0))},\ \|u\|_{W^{2,1}_\infty(\mathbb{R}^n\times[0,T])} \leq \Lambda, \\ \Lambda^{-1} \leq \beta(p),\ A(p),\ B(p) \leq \Lambda \quad \text{for } |p| = 1, \quad \text{and} \\ \Lambda^{-1} I \leq D^2 A. \end{gathered} \tag{8}$$

Define $F \in C^0(\mathbb{R}^n \times [0,T] \times (\mathbb{R}^n - \{0\}) \times S(n))$, where $S(n)$ denotes the set of all real, symmetric $n \times n$ matrices endowed with the usual ordering by

$$F(x,t,p,X) := -\beta(p)^{-1}\mathrm{tr}(B(p)D^2B(p)X) - \beta(p)^{-1}B(p)u(x,t). \tag{9}$$

Since B is homogeneous of degree one, it follows that F is geometric in the sense that

$$F(x,t,\lambda p, \lambda X + \sigma(p \otimes p)) = \lambda F(x,t,p,X) \quad \text{for } \lambda > 0,\ \sigma \in \mathbb{R},$$

where $\otimes$ denotes the tensor product $p \otimes q = (p_i q_j)_{i,j}$ on $\mathbb{R}^n$.

Our limit problem, the fully anisotropic mean-curvature flow, is

$$\beta(\nabla\omega)\partial_t\omega - \mathrm{tr}(B(\nabla\omega)D^2B(\nabla\omega)D^2\omega) - B(\nabla\omega)u = 0, \tag{10}$$

with the evolving interface given by $\Gamma_t := [\omega(.,t) = 0]$. Equation (10) was treated by Chen, Giga, and Goto in [7] when u was independent of space, but all of their methods apply to general u. In order to give meaning to (10), we seek a solution in the viscosity sense of

$$\partial_t\omega + F(x,t,\nabla\omega, D^2\omega) = 0.$$

Existence and comparison for (10) were proved in [7].

We recall the definition of viscosity solutions. In the following, we denote by LSC(...) and USC(...) the sets of lower semicontinuous and upper semicontinuous functions.

Definition 2.1. *Let $K \in C^0(\mathbb{R}^n \times [0,T] \times \mathbb{R} \times \mathbb{R} \times (\mathbb{R}^n - \{0\}) \times S(n))$. A function $v : Q \to \mathbb{R}$, where $\emptyset \neq Q \subseteq \mathbb{R}^n \times\ ]0,T[$ is open, is called a* viscosity subsolution *of*

$$K(x,t,v,\partial_t v, \nabla v, D^2 v) = 0 \quad \text{in } Q, \tag{11}$$

written as

$$K(x,t,v,\partial_t v, \nabla v, D^2 v) \leq 0 \quad \text{in } Q,$$

if $v \in \mathrm{USC}(Q)$ *and*

$$\forall (a,p,X) \in \mathrm{P}^{2,+}v(x,t),\ (x,t) \in Q: \qquad K_\star(x,t,v(x,t),a,p,X) \leq 0,$$

where $K_\star$ *is the* lower semicontinuous envelope of K, *that is,*

$$K_\star(x,t,r,a,p,X) \\ := \inf\{\liminf_{j\to\infty} K(x_j,t_j,r_j,a_j,p_j,X_j) \mid (x_j,t_j,r_j,a_j,p_j,X_j) \to (x,t,r,a,p,X)\}.$$

$\mathrm{P}^{2,+}$ *is the* set of superdifferentials *and is defined in the next subsection. A* supersolution *is defined analogously by considering* $K^\star$, *the upper semicontinuous envelope of* K, *and the set of subdifferentials* $\mathrm{P}^{2,-}$. v *is a* solution of (11) *when it is both a sub- and a supersolution. It is necessary to introduce the semicontinuous envelopes since* K *is not continuous when* $p = 0$.

The double-obstacle problem corresponding to (11) *is given by*

$$\max(v-1, \min(v+1, K(x,t,v,\partial_t v, \nabla v, D^2 v))) = 0 \quad \text{in } Q. \tag{12}$$

Equivalently, v *is a* subsolution of (12) *if*

$$v \in \mathrm{USC}(Q), \\ v \leq 1,$$

and

$$\forall (a,p,X) \in \mathrm{P}^{2,+}v(x,t),\ (x,t) \in Q,\ \text{with } v(x,t) > -1: \\ K_\star(x,t,v(x,t),a,p,X) \leq 0.$$

In this article, we mainly study parabolic equations and their double-obstacle problems, that is, where K is given by

$$K(x,t,v,\partial_t v, \nabla v, D^2 v) = \partial_t v + H(x,t,v,\nabla v, D^2 v).$$

Definition 2.2. *For* $v : Q \to \mathbb{R}$, *where* $\emptyset \neq Q \subseteq \mathbb{R}^n \times]0,T[$ *is open,* $(x_0,t_0) \in Q$, *we define the* sets of superdifferentials

$$\mathrm{P}^{2,+}v(x_0,t_0) := \{(a,p,X) \in \mathbb{R} \times \mathbb{R}^n \times S(n) \mid v(x,t) \leq v(x_0,t_0) \\ + a(t-t_0) + p(x-x_0) + \tfrac{1}{2}(x-x_0)^T X (x-x_0) \\ + o(|t-t_0| + |x-x_0|^2) \text{ as } t \to t_0,\ x \to x_0\},$$

and

$$\overline{\mathrm{P}}^{2,+}v(x_0,t_0) := \{(a,p,X) \in \mathbb{R} \times \mathbb{R}^n \times S(n) \mid \exists (a_j,p_j,X_j) \in \mathrm{P}^{2,+}v(x_j,t_j) : \\ (a_j,p_j,X_j) \to (a,p,X),\ (x_j,t_j) \to (x_0,t_0), v(x_j,t_j) \to v(x_0,t_0)\}.$$

The sets of subdifferentials $\mathrm{P}^{2,-}$ *and* $\overline{\mathrm{P}}^{2,-}$ *are defined analogously.*

Remark 2.3. It is seen easily that for $\varphi \in C^{2,1}(Q)$ with $v - \varphi \leq (v-\varphi)(x_0,t_0)$ in Q, the triple of derivatives $(\partial_t\varphi, \nabla\varphi, D^2\varphi)(x_0,t_0) \in \mathrm{P}^{2,+}v(x_0,t_0)$.

Conversely, for all superdifferentials $(a,p,X) \in \mathrm{P}^{2,+}v(x_0,t_0)$, there is a $\varphi \in C^{2,1}(Q)$ with $v - \varphi \leq (v-\varphi)(x_0,t_0)$ in Q and $(a,p,X) = (\partial_t\varphi, \nabla\varphi, D^2\varphi)(x_0,t_0)$. A proof of the second statement can be found in [19, section 14A].

2.2. The equation. We consider the anisotropic Allen–Cahn equation

$$\varepsilon\beta(\nabla\varphi)\partial_t\varphi - \varepsilon\nabla A'(\nabla\varphi) + \frac{1}{\varepsilon}W'(\varphi) - c_W u = 0 \quad \text{in } \mathbb{R}^n \times [0,T], \tag{13}$$

where $W(t) := (t^2-1)^2$. As already pointed out in the introduction, it is not clear how to define a weak solution of (13). Therefore, we treat (13) and its double-obstacle variant in the viscosity sense. To be precise, we define $G_\varepsilon,\ \tilde{G}_\varepsilon \in C^0(\mathbb{R}^n \times [0,T] \times \mathbb{R} \times (\mathbb{R}^n - \{0\}) \times S(n))$ by

(14)

$$G_\varepsilon(x,t,r,p,X) := -\varepsilon\beta(p)^{-1}\mathrm{tr}(D^2A(p)X) - \frac{1}{\varepsilon}\beta(p)^{-1}r - \beta(p)^{-1}\frac{\pi}{4}u(x,t)$$

and

$$\tilde{G}_\varepsilon(x,t,r,p,X) := -\varepsilon\beta(p)^{-1}\mathrm{tr}(D^2A(p)X) + \frac{1}{\varepsilon}\beta(p)^{-1}W'(r) - \beta(p)^{-1}c_W u(x,t).$$

Then in the viscosity formulation, the Allen–Cahn equation (13) reads

$$\partial_t\varphi + \frac{1}{\varepsilon}\tilde{G}_\varepsilon(.,.,\varphi,\nabla\varphi,D^2\varphi) = 0, \tag{15}$$

and its double-obstacle variant is given by

$$\max\left(\varphi - 1, \min\left(\varphi + 1, \partial_t\varphi + \frac{1}{\varepsilon}G_\varepsilon(.,.,\varphi,\nabla\varphi,D^2\varphi)\right)\right) = 0. \tag{16}$$

As already mentioned, existence and comparison for (10) were proved by Chen, Giga, and Goto in [7]. We will state these theorems without proof.

Throughout this article, we will compute on the whole of $\mathbb{R}^n$ and we will consider only space-periodic functions; therefore, we assume u and the initial data to be periodic in space. Here and in the following, *periodic* means *space periodic* and that there are n linearly independent periods.

THEOREM 2.4. *Let* $\emptyset \neq \Omega \subseteq \mathbb{R}^n$ *be open,* $0 < T \leq \Lambda,$ *and* $H \in C^0(\Omega \times [0,T] \times \mathbb{R} \times (\mathbb{R}^n - \{0\}) \times S(n))$ *satisfy*

(i) $|H(x,t,r,p,X)| \leq C(\Gamma)$ *for* $|r|,|p|,\|X\| \leq \Gamma,$

(ii) $H(x,t,r,p,X) \geq H(x,t,r,p,Y)$ *when* $X \leq Y,$

(iii) $H(x,t,r,p,X) - H(x,t,s,p,X) \geq -\Lambda(r-s)$ *when* $r \geq s,$

(iv) $|H(x,t,r,p,X) - H(y,t,r,p,X)| \leq \Lambda|x-y|(1+|p|),$ *and*

(v) $H_\star(x,t,r,0,0) = H^\star(x,t,r,0,0).$

Further, let $v \in \mathrm{USC}(\bar{\Omega} \times [0,T])$ *and* $w \in \mathrm{LSC}(\bar{\Omega} \times [0,T])$ *satisfy*

$$\partial_t v + H_\star(.,.,v,\nabla v,D^2 v) \leq 0,$$

and

$$\partial_t w + H^\star(.,.,w,\nabla w,D^2 w) \geq 0,$$

either

(a) *in* $\Omega \times]0,T[$

or

(b) *in* $\Omega \times]0,T[$ *in the double-obstacle sense.*

Next, we assume that either

(α) $\Omega = \mathbb{R}^n$ *and* v *and* w *are periodic with the same period*

or

(β) $\Omega \subset\subset \mathbb{R}^n$ *and* $v \leq w$ *on* $\partial\Omega \times [0, T[$.

Then

$$v(.,0) \leq w(.,0) \quad \textit{in } \Omega$$

implies

$$v \leq w \quad \textit{in } \Omega \times [0, T[.$$

Proof. This comparison principle was proved by Chen, Giga, and Goto in [7]. Properly, they considered H to be independent of x and they did not consider the double-obstacle problem. However, their proof applies to the general case. See also Theorem 2.12, where the ideas of [7] are applied to prove a modified comparison principle for the double-obstacle Allen–Cahn equation. ∎

THEOREM 2.5. *For periodic, continuous initial data,* (10) *has a unique solution.*

Proof. We again refer to [7]. ∎

Remark 2.6. When β is constant, one can easily check that G_ε and $\tilde{G}_\varepsilon$ of (14) satisfy the assumptions for H in Theorem 2.4. Therefore, the corresponding parabolic equations admit a comparison principle when they are considered in the viscosity sense. For nonconstant β, a comparison principle of this kind does not hold. This can be seen as follows. We take $u = 0$ and consider solutions which are constant in space. Such solutions w satisfy the ordinary differential equation

$$\varepsilon\beta^\star(t)w'(t) - \frac{1}{\varepsilon}w(t) = 0$$

and the inequality $|w(t)| < 1$, where $\beta^\star$ is any function satisfying $\inf\beta \leq \beta^\star \leq \sup\beta$. Taking $\beta^\star = \inf\beta$ and $\beta^\star = \sup\beta$, comparison is easily contradicted. Since (15) does not admit a comparison principle, the existence of solutions to (15) cannot be proved by Perron's method.

The rest of this section is devoted to proving existence of solutions to (15) and a modified comparison principle in which the sub- or supersolution in Theorem 2.4 satisfies additional conditions.

Existence is proved by approximating (15) through regularized equations. We will establish a uniform bound on the Hölder continuity of solutions of these regularized equations, hence getting a solution of (15). To carry out the limit procedure, we recall a definition of [9, section 6], which is stated below in Definition 2.7.

In section 4, we will construct sub- and supersolutions which admit the additional condition required in the modified comparison principle (Theorem 2.12). Therefore, we will be able to compare these sub- and supersolutions with a solution and conclude the desired convergence as $\varepsilon \to 0$ for these solutions.

DEFINITION 2.7. *For a family of functions* $v_\delta : Q \subseteq \mathbb{R}^N \to \mathbb{R}$, *we define*

$$v := \lim_{\delta \to 0_\star} v_\delta$$

by

$$v(z_0) := \inf\left\{\liminf_{j\to\infty} v_{\delta_j}(z_j) \mid \delta_j \to 0,\ z_j \to z_0\right\}.$$

$\lim_{\delta\to 0}{}^\star v_\delta$ *is defined analogously.*

2.3. Regularization. First, we consider (13) when $\beta \in C^\infty(\mathbb{R}^n),\ A \in C^\infty(\mathbb{R}^n)$, W' is replaced by any $f \in C^\infty(\mathbb{R}),$ and $u \in C^\infty(\mathbb{R}^n \times [0,T])$ is periodic in space. We drop the homogeneity assumptions on β and A, but we demand

$$\begin{aligned} \Lambda^{-1} &\leq \beta \leq \Lambda, \\ \Lambda^{-1} I &\leq D^2 A, \\ \Lambda^{-1}|p|^2 - \Lambda &\leq A(p), \\ \|u\|_{L^\infty(\mathbb{R}^n \times [0,T])},\ \|\nabla u\|_{L^\infty(\mathbb{R}^n \times [0,T])},\ |f(0)| &\leq \Lambda, \quad \text{and} \\ f' &\geq -\Lambda. \end{aligned} \tag{17}$$

We consider the equation

(18)

$$\partial_t \varphi - \beta(\nabla\varphi)^{-1} \operatorname{tr}(D^2 A(\nabla\varphi) D^2 \varphi) + \beta(\nabla\varphi)^{-1} f(\varphi) - \beta(\nabla\varphi)^{-1} u = 0 \quad \text{in } \mathbb{R}^n \times [0,T]$$

with periodic initial data $\varphi_0 \in C^\infty(\mathbb{R}^n)$ and

$$\|\varphi_0\|_{C^{0,1}(\mathbb{R}^n)} \leq \Lambda. \tag{19}$$

Since (18) admits a comparison principle, we conclude from (19) that a solution of (18) satisfies

$$\|\varphi\|_{L^\infty(\mathbb{R}^n \times [0,T])},\ \|\nabla\varphi\|_{L^\infty(\mathbb{R}^n \times [0,T])} \leq C(\Lambda). \tag{20}$$

With this a priori bound, we find using techniques of [14] that (18) has a unique, periodic solution $\varphi \in C^\infty(\mathbb{R}^n \times [0,T])$.

We write (18) in divergence form as

$$\beta(\nabla\varphi)\partial_t \varphi - \nabla A'(\nabla\varphi) + f(\varphi) - u = 0,$$

multiply by $\partial_t \varphi$, integrate over $K_{t_0} := K \times [0, t_0]$, where K is a periodic cell, and get

$$\int_{K_{t_0}} \beta(\nabla\varphi)|\partial_t\varphi|^2 + A'(\nabla\varphi)\partial_t\nabla\varphi + f(\varphi)\partial_t\varphi = \int_{K_{t_0}} u\partial_t\varphi. \tag{21}$$

We define $F(t) := \int_0^t (f(s) - f(0) + (\Lambda+1)s)ds$ and get $F'(t) = f(t) - f(0) + (\Lambda+1)t,\ F'(0) = 0,$ and $F'' \geq 1$. From (20) and (21), we conclude

$$\begin{aligned} \Lambda^{-1} \int_{K_{t_0}} |\partial_t\varphi|^2 &+ \int_K A(\nabla\varphi(t_0)) + F(\varphi(t_0)) \\ &\leq \int_K A(\varphi_0) + F(\varphi_0) + \int_{K_{t_0}} ((\Lambda+1)\varphi - f(0) + u)\partial_t\varphi \\ &\leq \int_K A(\nabla\varphi_0) + F(\varphi_0) + C(\Lambda)|K| + \frac{1}{2\Lambda}\int_{K_{t_0}} |\partial_t\varphi|^2. \end{aligned}$$

Assuming

$$\int_K A(\nabla\varphi_0) + F(\varphi_0) \leq \Lambda \tag{22}$$

we obtain, noting (17),

$$\begin{aligned} &\|\partial_t\varphi\|_{L^2(0,T;L^2(K))},\ \|A(\nabla\varphi)\|_{L^\infty(0,T;L^2(K))}, \\ &\|F(\varphi)\|_{L^\infty(0,T;L^1(K))},\ \|\varphi\|_{L^2(0,T;H^{1,2}(K))} \leq C(\Lambda). \end{aligned} \tag{23}$$

As in [11], the above $L^2(0,T;L^2(K))$ estimate on $\partial_t\varphi$ together with (20) yields

$$\|\varphi\|_{H^{1/n+1,1/2(n+1)}(\mathbb{R}^n\times[0,T])} \leq C(\Lambda), \tag{24}$$

where $H^{\alpha,\alpha/2}$ denotes the space of functions that are Hölder continuous of exponent α in space and exponent $\frac{\alpha}{2}$ in time.

Now we approximate β, A, and u of section 2.1 appropriately by β_δ, A_δ, $u_\delta \in \mathrm{C}^\infty$ as in (17). In the case of the smooth well, we take $f_\delta = W'$; for the double-obstacle problem, we take $f_\delta := g' + \frac{1}{\delta}h'$ with $g \geq 0$, $g'(t) = -t$ for $|t| \leq 1$, $|g''| \leq \Lambda$, $h(t) = 0$ for $|t| \leq 1$, $th'(t) > 0$ for $|t| > 1$, and $0 \leq h'' \leq \Lambda$. Taking the approximation such that $A_\delta \to A$ uniformly on compact subsets of $\mathbb{R}^n$, we have in the smooth case

$$\int_K A_\delta(\nabla\varphi_0) + F_\delta(\varphi_0) \leq C(\Lambda) \tag{25}$$

for small δ. Since $|\varphi_0| \leq 1$ is required in the double-obstacle problem, we have

$$\int_K \left(g + \frac{1}{\delta}h\right)(\varphi_0) = \int_K g(\varphi_0) \leq C(g)|K|,$$

which yields (25) as well. From (22) and (23), we conclude that (24) is satisfied for the unique solution φ_δ of (18) when (β, A, f, u) is substituted by $(\beta_\delta, A_\delta, f_\delta, u_\delta)$. Equation (18) can be considered as a viscosity equation of the form

$$\partial_t\varphi_\delta + H_\delta(.,.,\varphi_\delta, \nabla\varphi_\delta, D^2\varphi_\delta) = 0 \quad \text{in } \mathbb{R}^n \times]0,T[.$$

Choosing β_δ and A_δ as convolutions—that is, $\beta_\delta(p) := \int_{\mathbb{R}} \beta(q)\eta_\delta(p-q)dq$ and $A_\delta(p) := \int_{\mathbb{R}} A(q)\eta_\delta(p-q)dq$, where $\eta_\delta(p) := \delta^{-1}\eta(\frac{p}{\delta})$ with $\eta \in C^\infty(\mathbb{R})$, $\eta \geq 0$, and $\int \eta = 1$—it is easily seen that

$$\lim_{\delta\to 0_\star}{}^{(\star)} (\beta_\delta(p)(a + H_\delta(x,t,r,p,X))) = (\beta(p)(a + \tilde{G}_1(x,t,r,p,X)))_\star^{(\star)}$$

when the smooth well is considered and that

$$\begin{aligned} &\lim_{\delta\to 0_\star} (a + H_\delta(x,t,r,p,X)) \leq 0 \\ &\qquad \Rightarrow \max(r-1, \min(r+1, a + G_1(x,t,r,p,X)))_\star \leq 0 \end{aligned}$$

and

$$\begin{aligned} &\lim_{\delta\to 0}{}^\star (a + H_\delta(x,t,r,p,X)) \geq 0 \\ &\qquad \Rightarrow \max(r-1, \min(r+1, a + G_1(x,t,r,p,X)))^\star \geq 0 \end{aligned}$$

when the double-obstacle problem is considered. Taking a uniformly convergent subsequence, passing to the limits, and applying [9, Lemma 6.1], we obtain a solution of (15) and its double-obstacle variant. Since only an estimate of $\|\varphi_0\|_{C^{0,1}(\mathbb{R}^n)}$ was required, we have proved the following existence theorem.

Theorem 2.8. *For every periodic* $\varphi_0 \in C^{0,1}(\mathbb{R}^n)$*, there are periodic viscosity solutions* $\varphi, \tilde{\varphi} \in H^{1/(n+1),1/2(n+1)}(\mathbb{R}^n \times [0,T])$ *to the anisotropic double-obstacle Allen–Cahn and its smooth form, that is,*

$$\max\left(\varphi - 1, \min\left(\varphi + 1, \partial_t\varphi + \frac{1}{\varepsilon}G_\varepsilon(.,.,\varphi,\nabla\varphi,D^2\varphi)\right)\right) = 0 \quad \textit{in } \mathbb{R}^n\times\ ]0,T[,$$

$$\partial_t\tilde{\varphi} + \frac{1}{\varepsilon}\tilde{G}_\varepsilon(.,.,\tilde{\varphi},\nabla\tilde{\varphi},D^2\tilde{\varphi}) = 0 \quad \textit{in } \mathbb{R}^n \times]0,T[,$$

and

$$\varphi(.,0) = \tilde{\varphi}(.,0) = \varphi_0 \quad \textit{in } \mathbb{R}^n.$$

Here G_ε and $\tilde{G}_\varepsilon$ are defined in (14).

Remark 2.9. Taking $H = \frac{1}{\varepsilon} G_\varepsilon$ in Theorem 2.4, we observe that it satisfies all conditions except

$$\text{(v)} \quad G_{\varepsilon,\star}(x,t,r,0,0) = G^\star_\varepsilon(x,t,r,0,0).$$

In the next theorem, we impose an additional condition on the sub- or supersolution of Theorem 2.4 for $H = \frac{1}{\varepsilon} G_\varepsilon$ which is sufficient to establish comparison. To prove this theorem, we apply the following theorem, which is given in a more general version in [9, Theorem 8.3]. Since we apply it to the double-obstacle problem, we must state it with one detail slightly changed.

Theorem 2.10. *Let $v \in \mathrm{USC}(\mathbb{R}^n \times\,]0,T[)$ and $w \in \mathrm{LSC}(\mathbb{R}^n \times\,]0,T[)$, and define $\Phi(t,x,y) := v(x,t) - w(y,t) - \alpha|x-y|^2$, where $\alpha \geq 0$. We suppose that $\Phi(t_0,x_0,y_0) = \sup_{]0,T[\,\times \mathbb{R}^n \times \mathbb{R}^n} \Phi$, where $(t_0,x_0,y_0) \in\,]0,T[\,\times \mathbb{R}^n \times \mathbb{R}^n$. We assume that there is an $r > 0$ such that for $\Gamma > 0$, $(a,p,X) \in \mathrm{P}^{2,+}v(x,t)$,*

$$|x - x_0| + |t - t_0| + |v(x,t) - v(x_0,t_0)| \leq r \quad \textit{and} \quad |p| + \|X\| \leq \Gamma$$
$$\textit{implies } a \leq C(\Gamma);$$

likewise, for $(a,p,X) \in \mathrm{P}^{2,-}w(y,t)$,

$$|y - y_0| + |t - t_0| + |w(y,t) - w(y_0,t_0)| \leq r \quad \textit{and} \quad |p| + \|X\| \leq \Gamma$$
$$\textit{implies } a \geq -C(\Gamma).$$

Then there are $a \in \mathbb{R}$ and $X, Y \in S(n)$ such that

$$(a, 2\alpha(x_0 - y_0), X) \in \overline{\mathrm{P}}^{2,+}v(x_0,t_0),$$
$$(a, 2\alpha(x_0 - y_0), Y) \in \overline{\mathrm{P}}^{2,-}w(y_0,t_0),$$

and

$$X \leq Y.$$

Moreover, when $\alpha = 0$, we get

$$X \leq 0 \leq Y.$$

Remark 2.11. We observe that the condition about the boundedness of the time derivative of v is satisfied if v is a subsolution of a parabolic equation. In this case, it is indeed sufficient to require merely that $|v(x,t)| \leq \Gamma$, as is done in [9, Theorem 8.3]. If v is a solution of a double-obstacle problem, this bound can only be concluded if $v(x,t) > -1$. Therefore, we have to replace $|v(x,t)| \leq \Gamma$ by $|v(x,t) - v(x_0,t_0)| \leq r$ since we will apply the theorem when $v(x_0,t_0) > -1$.

Theorem 2.12. *We assume that v and w are periodic with the same period, that they are sub- and supersolutions, respectively, of*

$$\max\left(\varphi - 1, \min\left(\varphi + 1, \partial_t \varphi + \frac{1}{\varepsilon} G_\varepsilon(.,.,\varphi,\nabla\varphi,D^2\varphi)\right)\right) = 0 \quad \textit{in } \mathbb{R}^n \times\,]0,T[,$$

and that $v(.,0) \leq w(.,0)$. Moreover, we assume that for $(a,0,X) \in \overline{\mathrm{P}}^{2,-}w(x_0,t_0)$ with $(x_0,t_0) \in \mathbb{R}^n \times\,]0,T[$ and $w(x_0,t_0) < 1$,

$$(26) \qquad \limsup_{p \to 0,\ p \neq 0} \left(\varepsilon\beta^\star a - \varepsilon tr(D^2A(p)X) - \frac{1}{\varepsilon} w(x_0,t_0) - \frac{\pi}{4} u(x_0,t_0)\right) \geq 0$$

holds for any $\inf\beta \le \beta^\star \le \sup\beta$*. Then the following comparison inequality holds:*

$$v \le w \quad \textit{in } \mathbb{R}^n \times [0,T[.$$

Proof. We prove the theorem by contradiction. Suppose there is $(\bar{x},\bar{t}) \in \mathbb{R}^n \times]0,T[$ such that $v(\bar{x},\bar{t}) > w(\bar{x},\bar{t})$. Choose T_0 so that $0 < \bar{t} < T_0 < T$. We know that v and w are upper and lower semicontinuous, respectively, so periodicity implies

$$\sup_{y\in\mathbb{R}^n} (v(y,T_0) - w(y,T_0)) < \infty;$$

hence for λ large enough, we have

$$\text{(27)} \qquad \forall y \in \mathbb{R}^n\colon\ \exp(-\lambda t)(v(x,t) - w(x,t)) > \exp(-\lambda T_0)(v(y,T_0) - w(y,T_0)).$$

We assume additionally that $\lambda > C(\Lambda)\varepsilon^{-2}$.

We define

$$\tilde{v}(x,t) := \exp(-\lambda t)v(x,t)$$

and

$$\tilde{w}(x,t) := \exp(-\lambda t)w(x,t).$$

We conclude from (27) that

$$\sup_{\mathbb{R}^n\times[0,T_0]} (\tilde{v} - \tilde{w}) =: \delta > 0,$$

and, since $\tilde{v}$ and $\tilde{w}$ are upper and lower semicontinuous, respectively, that there exists $(x_0,t_0) \in \mathbb{R}^n \times [0,T_0]$ such that

$$\text{(28)} \qquad \tilde{v}(x_0,t_0) - \tilde{w}(x_0,t_0) = \sup_{\mathbb{R}^n\times[0,T_0]} (\tilde{v} - \tilde{w}),$$

and, noting that $(\tilde{v} - \tilde{w})(.,0) \le 0$, for any such (x_0,t_0),

$$\text{(29)} \qquad 0 < t_0 < T_0.$$

For proving comparison for viscosity solutions, we proceed with the standard technique of doubling the number of space variables and penalizing this doubling. We define for $\alpha > 0$ and $\eta \in \mathbb{R}^n$ the upper semicontinuous functions

$$\text{(30)} \qquad \begin{aligned} \Psi(t,x,y) &:= \tilde{v}(x,t) - \tilde{w}(y,t), \\ \Phi_\alpha(t,x,y) &:= \Psi(t,x,y) - \alpha|x-y|^2, \quad \text{and} \\ \Phi_{\alpha,\eta}(t,x,y) &:= \Psi(t,x,y) - \alpha|x-y-\eta|^2. \end{aligned}$$

We distinguish two cases and in each derive a contradiction.

Case (i). There is a $\mu > 0$ such that for all $|\eta| < \mu$ one of the maximum points (t_η, x_η, y_η) of $\Phi_{\alpha,\eta}$, that is

$$\Phi_{\alpha,\eta} \le \Phi_{\alpha,\eta}(t_\eta,x_\eta,y_\eta) \quad \text{in } [0,T_0]\times\mathbb{R}^n\times\mathbb{R}^n,$$

satisfies

$$\text{(31)} \qquad x_\eta - y_\eta = \eta.$$

We define $f(\eta) := \sup_{x-y=\eta} \Psi$ and get

$$f(\eta) \geq \Psi(t_\eta, x_\eta, y_\eta) = \Phi_{\alpha,\eta}(t_\eta, x_\eta, y_\eta) \geq \sup_{x-y=\xi} \Phi_{\alpha,\eta} = f(\xi) - \alpha|\xi - \eta|^2$$

for all $\xi,\ \eta \in U_\mu(0)$. Therefore, $f : U_\mu(0) \to \mathbb{R}$ is constant and

$$\sup_{x-y=\eta} \Psi = f(\eta) = f(0) = \sup_{x=y} \Psi = \sup_{\mathbb{R}^n \times [0,T_0]} (\tilde{v} - \tilde{w}) = \tilde{v}(x_0, t_0) - \tilde{w}(x_0, t_0)$$

by (28). This yields

$$\forall x, y \in U_{\mu/2}(x_0)\colon \quad \forall t \in [0, T_0]\colon \quad \tilde{v}(x,t) - \tilde{w}(y,t) \leq \tilde{v}(x_0,t_0) - \tilde{w}(x_0,t_0).$$

Applying Theorem 2.10 to $\Phi = \Psi$, we get

$$\begin{aligned} \exists \tilde{a},\ \tilde{X},\ \tilde{Y}\colon \quad & (\tilde{a}, 0, \tilde{X}) \in \overline{\mathrm{P}}^{2,+} v(x_0, t_0), \\ & (\tilde{a}, 0, \tilde{Y}) \in \overline{\mathrm{P}}^{2,-} w(x_0, t_0), \\ & \tilde{X} \leq 0 \leq \tilde{Y}. \end{aligned}$$

Setting $a := \tilde{a} \exp(\lambda t_0)$, $X := \tilde{X} \exp(\lambda t_0)$, and $Y = \tilde{Y} \exp(\lambda t_0)$ and noting that $\exp(\lambda(t - t_0)) v(x_0, t_0) = (1 + \lambda(t - t_0)) v(x_0, t_0) + O(|t - t_0|^2)$, we have

$$\begin{aligned} (a + \lambda v(x_0, t_0), 0, X) &\in \overline{\mathrm{P}}^{2,+} v(x_0, t_0), \\ (a + \lambda w(x_0, t_0), 0, Y) &\in \overline{\mathrm{P}}^{2,-} w(x_0, t_0), \end{aligned}$$

and

$$X \leq 0 \leq Y.$$

Furthermore, the inequality $1 \geq v(x_0, t_0) > w(x_0, t_0) \geq -1$ holds. Since v and w are sub- and supersolutions, respectively, we obtain

$$\begin{aligned} a + \lambda v(x_0, t_0) + \frac{1}{\varepsilon} G_{\varepsilon,\star}(x_0, t_0, v(x_0, t_0), 0, X) \leq 0 \quad \text{and} \\ a + \lambda w(x_0, t_0) + \frac{1}{\varepsilon} G_{\varepsilon}^{\star}(x_0, t_0, w(x_0, t_0), 0, Y) \geq 0. \end{aligned} \tag{32}$$

Using the definition of the semicontinuous envelope, we conclude from (32) that

$$\begin{aligned} 0 &\geq \liminf_{p \to 0,\ p \neq 0} \left(\varepsilon \beta(p)(a + \lambda v(x_0, t_0)) - \varepsilon \operatorname{tr}(D^2 A(p) X) - \frac{1}{\varepsilon} v(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \right) \\ &\geq \liminf_{p \to 0,\ p \neq 0} \left(\varepsilon \beta(p)(a + \lambda v(x_0, t_0)) - \frac{1}{\varepsilon} v(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \right) \end{aligned}$$

since $X \leq 0$. Therefore,

$$0 \geq \varepsilon \beta^\star (a + \lambda v(x_0, t_0)) - \frac{1}{\varepsilon} v(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \tag{33}$$

for some $\inf \beta \leq \beta^\star \leq \sup \beta$. Taking this $\beta^\star$ and $(\tilde{a}, 0, \tilde{Y}) \in \overline{\mathrm{P}}^{2,-} w(x_0, t_0)$ in (26),

(34)

$$\begin{aligned} 0 &\leq \limsup_{p \to 0,\ p \neq 0} (\varepsilon \beta^\star (a + \lambda w(x_0, t_0)) - \varepsilon \operatorname{tr} \left(D^2 A(p) Y) - \frac{1}{\varepsilon} w(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \right) \\ &\leq \varepsilon \beta^\star (a + \lambda w(x_0, t_0)) - \frac{1}{\varepsilon} w(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \end{aligned}$$

since $Y \geq 0$. From (33) and (34), we obtain

$$0 \geq \left(\varepsilon\beta^\star\lambda - \frac{1}{\varepsilon}\right)(v(x_0,t_0) - w(x_0,t_0)) > 0,$$

which is a contradiction.

Case (ii) For all $\mu > 0$, there is an $|\eta| < \mu$ such that one of the maximum points (t_η, x_η, y_η) of $\Phi_{\alpha,\eta}$, that is,

$$\Phi_{\alpha,\eta} \leq \Phi_{\alpha,\eta}(t_\eta, x_\eta, y_\eta) \quad \text{in } [0,T_0] \times \mathbb{R}^n \times \mathbb{R}^n,$$

satisfies

$$x_\eta - y_\eta \neq \eta. \tag{35}$$

From periodicity and upper semicontinuity, we obtain for a subsequence

$$\begin{aligned} (t_\eta, x_\eta, y_\eta) \to (t_\alpha, x_\alpha, y_\alpha) \in [0,T_0] \times \mathbb{R}^n \times \mathbb{R}^n, \quad \Phi_\alpha \leq \Phi_\alpha(t_\alpha, x_\alpha, y_\alpha) \quad \text{and} \\ \tilde{v}(x_\eta, t_\eta) - \tilde{w}(y_\eta, t_\eta) \to \tilde{v}(x_\alpha, t_\alpha) - \tilde{w}(y_\alpha . t_\alpha). \end{aligned} \tag{36}$$

From [9, Proposition 3.7], for a subsequence $\alpha \to \infty$, we get

$$\begin{aligned} (t_\alpha, x_\alpha, y_\alpha) &\to (t_0, x_0, y_0) \in [0,T_0] \times \mathbb{R}^n \times \mathbb{R}^n, \quad x_0 = y_0, \\ \tilde{v}(x_\alpha, t_\alpha) - \tilde{w}(y_\alpha, t_\alpha) &\to \tilde{v}(x_0,t_0) - \tilde{w}(x_0,t_0) = \sup_{\mathbb{R}^n \times [0,T_0]} (\tilde{v} - \tilde{w}). \end{aligned} \tag{37}$$

It follows from (28) that

$$\begin{aligned} & 0 < t_0 < T_0; \\ \text{hence} \quad & 0 < t_\alpha < T_0, \quad v(x_\alpha, t_\alpha) > w(y_\alpha, t_\alpha) \quad \text{for } \alpha \text{ large} \\ \text{and} \quad & 0 < t_\eta < T_0, \quad v(x_\eta, t_\eta) > w(y_\eta, t_\eta) \quad \text{for } |\eta| \text{ small.} \end{aligned}$$

Applying Theorem 2.10 to $\Phi = \Phi_{\alpha,\eta}(.,.,. - \eta)$ yields, after multiplying with $\exp(\lambda t_0)$,

$$\begin{aligned} \exists a,\ X,\ Y\colon \quad & (a + \lambda v(x_\eta, t_\eta), p_\eta, X) \in \overline{\mathrm{P}}^{2,+} v(x_\eta, t_\eta), \\ & (a + \lambda w(y_\eta, t_\eta), p_\eta, Y) \in \overline{\mathrm{P}}^{2,-} w(y_\eta, t_\eta), \\ & X \leq Y, \\ & p_\eta = 2\alpha(x_\eta - y_\eta - \eta)\exp(\lambda t_\eta). \end{aligned}$$

It also holds that $1 \geq v(x_\eta, t_\eta) > w(y_\eta, t_\eta) \geq -1$. Since v and w are sub- and supersolutions, we obtain

$$\begin{aligned} 0 &\geq \varepsilon(a + \lambda v(x_\eta, t_\eta)) + G_\varepsilon(x_\eta, t_\eta, v(x_\eta, t_\eta), p_\eta, X) \\ &\quad - \varepsilon(a + \lambda w(y_\eta, t_\eta)) - G_\varepsilon(y_\eta, t_\eta, w(y_\eta, t_\eta), p_\eta, Y) \\ &\geq \left(\varepsilon\lambda - C(\Lambda)\frac{1}{\varepsilon}\right)(v(x_\eta, t_\eta) - w(y_\eta, t_\eta)) - C(\Lambda)|u(x_\eta, t_\eta) - u(y_\eta, t_\eta)| \\ &\to \left(\varepsilon\lambda - C(\Lambda)\frac{1}{\varepsilon}\right)(v(x_0,t_0) - w(x_0,t_0)) > 0, \end{aligned}$$

which is again a contradiction.

Therefore, $v \leq w$ in $\mathbb{R}^n \times [0,T[$. □

Remark 2.13. The conclusion of the above proposition remains true if (26) is replaced by the analogous condition for v. More precisely, instead of (26), we require for $(a, 0, X) \in \overline{\mathrm{P}}^{2,+} v(x_0, t_0)$ with $(x_0, t_0) \in \mathbb{R}^n \times]0, T[$ and $v(x_0, t_0) > -1$ that

$$\liminf_{p \to 0,\ p \neq 0} \left(\varepsilon \beta^\star a - \varepsilon \mathrm{tr}(D^2 A(p) X) - \frac{1}{\varepsilon} v(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0) \right) \leq 0$$

for any $\inf \beta \leq \beta^\star \leq \sup \beta$.

This modified comparison principle is also valid for the smooth Allen–Cahn equation with obvious changes.

3. The distance function. As already pointed out in the introduction, we use a distance function that is induced by a Finsler metric for the construction of the sub- and supersolutions. This section is devoted to the presentation of some definitions and properties of the Finsler metric (see [4]) and the induced distance function. Most of these properties are known in the isotropic case; see [2]. However, Lemma 3.4 and the viscosity estimate of the time derivative of the distance function in Proposition 3.6 were not required in the isotropic case, but they will be used in section 4 for proving convergence in the anisotropic case.

3.1. The dual. We consider $\beta,\ A,\ B$, and u as in (8). As in [4], we define the dual of B,

$$B^\circ(q) := \sup\{ qp \mid B(p) \leq 1 \}. \tag{38}$$

B° is convex and homogeneous of degree one. Since $A \in \mathrm{C}^2(\mathbb{R}^{\mathrm{n}} - \{0\})$ is strictly convex, $B^\circ \in \mathrm{C}^2(\mathbb{R}^{\mathrm{n}} - \{0\})$ and satisfies

$$\begin{gathered} c_0(\Lambda) \leq B^\circ(q) \leq C(\Lambda) \quad \text{for } |q| = 1, \\ B(\nabla B^\circ(q)) = 1 \quad \text{for } q \neq 0, \\ \nabla B(\nabla B^\circ(q)) D^2 B^\circ(q) = 0 \quad \text{for } q \neq 0, \quad \text{and} \\ \|B^\circ\|_{C^2(B_2(0) - B_{1/2}(0))} \leq C(\Lambda). \end{gathered} \tag{39}$$

3.2. The distance function. We define a nonsymmetric metric d in $\mathbb{R}^n$ by

$$d(x, y) := B^\circ(x - y). \tag{40}$$

We easily obtain for $x, y, z \in \mathbb{R}^n$ that

$$\begin{gathered} d(x, z) \leq d(x, y) + d(y, z), \\ d(x + z, y + z) = d(x, y), \quad \text{and} \\ c_0(\Lambda)|x - y| \leq d(x, y) \leq C(\Lambda)|x - y|. \end{gathered} \tag{41}$$

Let $\omega \in \mathrm{LSC}(\mathbb{R}^n \times [0, T[)$ be a supersolution of

$$\begin{gathered} \beta(\nabla \omega) \partial_t \omega - \mathrm{tr}(B(\nabla \omega) D^2 B(\nabla \omega) D^2 \omega) - B(\nabla \omega) u \geq 0 \quad \text{in } \mathbb{R}^n \times]0, T[, \quad \text{or} \\ \partial_t \omega + F^\star(.,., \nabla \omega, D^2 \omega) \geq 0, \end{gathered} \tag{42}$$

where F is defined in (14). We define the distance function δ as follows:

$$\delta(x, t) := \inf_{y, \omega(y,t) \leq 0} d(x, y). \tag{43}$$

(41) implies

$$|\delta(x, t) - \delta(y, t)| \leq C(\Lambda)|x - y|. \tag{44}$$

We will prove some properties of δ; most of them are well known.

The next three lemmas were proved by Barles, Soner, and Souganidis in [2] in the isotropic case.

LEMMA 3.1. $\omega_\infty := \chi_{[\omega>0]}$ *is a supersolution of* (42).

Proof. From [7], we know that $\omega_\varepsilon := \min(1, \max(\frac{\omega}{\varepsilon}, 0))$ is a supersolution of (42). Because $\omega_\infty = \lim_{\varepsilon\to 0\star}\omega_\varepsilon$, defined in Definition 2.7, we get from [9, Lemma 6.1] that ω_∞ is a supersolution as well. □

LEMMA 3.2. *We define* $\delta^k(x,t) := \inf_{y\in\mathbb{R}^n}(k\omega_\infty(y,t) + d(x,y))$. *Then*

(i) $\delta^k = \min(\delta, k) \in \mathrm{LSC}(\mathbb{R}^n \times [0,T[)$ *and*

(ii) $\partial_t\delta^k + F^\star(.,.,\nabla\delta^k, D^2\delta^k) + C(\Lambda)|\nabla\delta^k|\delta^k \geq 0$ *in* $\mathbb{R}^n \times]0,T[$ *in the viscosity sense.*

Proof. (i) Trivially, we get

$$0 \leq \delta^k(x,t) \leq k\omega_\infty(x,t) + d(x,x) \leq k.$$

Second, if $\delta(x,t) = d(x,y)$ with $\omega(y,t) \leq 0$, then $\omega_\infty(y,t) = 0$ and

$$\delta^k(x,t) \leq d(x,y) = \delta(x,t);$$

hence

$$0 \leq \delta^k \leq \min(\delta, k).$$

Moreover, $\forall y \in \mathbb{R}^n : \ [\omega(y,t) \leq 0 \Rightarrow d(x,y) \geq \delta(x,t)]$. This yields

$$k\omega_\infty(y,t) + d(x,y) \geq \min(\delta(x,t), k);$$

hence

$$\delta^k \geq \min(\delta, k).$$

Now let $(x_j,t_j) \to (x_0,t_0) \in \mathbb{R}^n \times [0,T[$ and $\delta^k(x_j,t_j) = k\omega_\infty(y_j,t_j) + d(x_j,y_j)$. We obtain $d(x_j,y_j) \leq k$ and, for a subsequence, $y_j \to y_0$. This yields

$$\liminf_{j\to\infty} \delta^k(x_j,t_j) \geq k\omega_\infty(y_0,t_0) + d(x_0,y_0) \geq \delta^k(x_0,t_0).$$

(ii) Let $\varphi \in C^{2,1}(\mathbb{R}^n \times [0,T[)$ with $\delta^k - \varphi \geq (\delta^k - \varphi)(x_0,t_0),\ 0 < t_0 < T$, and $\delta^k(x_0,t_0) = k\omega_\infty(y_0,t_0) + d(x_0,y_0)$. Defining

$$\psi(y,t) := \varphi(y - y_0 + x_0, t), \quad \psi \in C^{2,1}(\mathbb{R}^n \times [0,T[),$$

we get, for $(y,t) \in \mathbb{R}^n \times]0,T[$ and $x := y - y_0 + x_0$,

$$\begin{aligned} k\omega_\infty(y,t) + d(x,y) - \psi(y,t) &\geq \delta^k(x,t) - \varphi(x,t) \\ &\geq \delta^k(x_0,t_0) - \varphi(x_0,t_0) = k\omega_\infty(y_0,t_0) + d(x_0,y_0) - \psi(y_0,t_0). \end{aligned} \tag{45}$$

Since $x - y = x_0 - y_0$, we get $d(x,y) = d(x_0,y_0)$, and (45) together with Lemma 3.1 yields

$$\partial_t\psi(y_0,t_0) + F^\star(y_0,t_0,\nabla\psi(y_0,t_0), D^2\psi(y_0,t_0)) \geq 0.$$

This implies

$$\begin{aligned}&\partial_t\varphi(x_0,t_0)+F^\star(x_0,t_0,\nabla\varphi(x_0,t_0),D^2\varphi(x_0,t_0))\\&\quad\geq F^\star(x_0,t_0,\nabla\varphi(x_0,t_0),D^2\varphi(x_0,t_0))-F^\star(y_0,t_0,\nabla\varphi(x_0,t_0),D^2\varphi(x_0,t_0))\\&\quad=\beta(\nabla\varphi(x_0,t_0))^{-1}B(\nabla\varphi(x_0,t_0))(u(y_0,t_0)-u(x_0,t_0))\\&\quad\geq -C(\Lambda)|\nabla\varphi(x_0,t_0)||x_0-y_0|\\&\quad\geq -C(\Lambda)|\nabla\varphi(x_0,t_0)|\delta^k(x_0,t_0)\end{aligned}$$

since $\delta^k(x_0,t_0)=k\omega_\infty(y_0,t_0)+d(x_0,y_0)\geq d(x_0,y_0)\geq c_0(\Lambda)|x_0-y_0|$. □

PROPOSITION 3.3. *δ is lower semicontinuous and is a viscosity supersolution of*

$$\partial_t\delta+F^\star(.,.,\nabla\delta,D^2\delta)+C(\Lambda)|\nabla\delta|\delta\geq 0\quad \textit{in } \mathbb{R}^n\times]0,T[.$$

Proof. Let $(x_j,t_j)\to(x_0,t_0)$ and choose $k>\delta(x_0,t_0)$. We obtain

$$\liminf_{j\to\infty}\delta(x_j,t_j)\geq\liminf_{j\to\infty}\delta^k(x_j,t_j)\geq\delta^k(x_0,t_0)=\delta(x_0,t_0),$$

where we have used the lower semicontinuity of δ^k, established in Lemma 3.2(i).

To prove that δ is a viscosity supersolution, we pass to the limit in Lemma 3.2(ii). According to [9, Lemma 6.1], viscosity supersolutions are preserved under the limit procedure defined in Definition 2.7. Hence it suffices to prove

$$\delta=\lim_{k\to\infty_\star}\delta^k. \tag{46}$$

From Lemma 3.2(i),

$$\delta=\lim_{k\to\infty}\delta^k\geq\lim_{k\to\infty_\star}\delta^k.$$

Conversely, let $(x_j,t_j)\to(x_0,t_0)$ and $k_j\to\infty$ with

$$\limsup_{j\to\infty}\delta^{k_j}(x_j,t_j)\leq\left(\lim_{k\to\infty_\star}\delta^k\right)(x_0,t_0)+\tau\leq\delta(x_0,t_0)+\tau$$

for some $\tau>0$. For j large, we get $\delta^{k_j}(x_j,t_j)<k_j$; hence

$$\delta(x_0,t_0)\leq\liminf_{j\to\infty}\delta(x_j,t_j)=\liminf_{j\to\infty}\delta^{k_j}(x_j,t_j)\leq\left(\lim_{k\to\infty_\star}\delta^k\right)(x_0,t_0)+\tau,$$

and (46) is established, concluding the proof. □

LEMMA 3.4. *For $x_0\in\mathbb{R}^n$ and $0\leq t_0\leq t_1<T$, the inequality*

$$\mu(\delta(x_0,t_1))\geq\mu(\delta(x_0,t_0))-C(\Lambda)(t_1-t_0) \tag{47}$$

holds, where $\mu(r):=\int_0^r\frac{s}{1+s}ds$.

Proof. The function μ, used below to define a subsolution for (51), appears in [7].

It suffices to prove the assertion when $\varrho:=\delta(x_0,t_0)>0$. We define $v(x,t):=\mu(\varrho)-\Gamma(t-t_0)-\mu(B^\circ(x_0-x))$ for some positive constant Γ chosen below, and we observe from Lemma 3.1 that $v\in C^{2,1}((\mathbb{R}^n-\{x_0\})\times[0,T[)$. For $x\neq x_0$, we get

$$\begin{aligned}\partial_tv(x,t)&=-\Gamma,\\\nabla v(x,t)&=\mu'(B^\circ(x_0-x))\nabla B^\circ(x_0-x)\\&=B^\circ(x_0-x)\nabla B^\circ(x_0-x)(1+B^\circ(x_0-x))^{-1},\quad\text{and}\\D^2v(x,t)&=-B^\circ(x_0-x)D^2B^\circ(x_0-x)(1+B^\circ(x_0-x))^{-1}\\&\quad-\nabla B^\circ(x_0-x)\otimes\nabla B^\circ(x_0-x)(1+B^\circ(x_0-x))^{-1}\\&\quad+B^\circ(x_0-x)\nabla B^\circ(x_0-x)\otimes\nabla B^\circ(x_0-x)(1+B^\circ(x_0-x))^{-2}.\end{aligned} \tag{48}$$

We observe that $v \in C^{1,1}(\mathbb{R}^n \times [0,T[)$ and satisfies

$$\begin{aligned} \nabla v(x_0,t) &= 0 \quad \text{and} \\ \|D^2 v(x,t)\| &\leq C(\Lambda)(1+B^\circ(x_0-x))^{-1}, \end{aligned} \tag{49}$$

where we have used (39). We conclude that for $x \neq x_0$,

$$\begin{aligned} &\partial_t v(x,t) + F(x,t,\nabla v(x,t), D^2 v(x,t)) \\ &\quad \leq -\Gamma + (1+B^\circ(x_0-x))^{-1} F(x,t,B^\circ(x_0-x)\nabla B^\circ(x_0-x), -C(\Lambda)I) \\ &\quad \leq -\Gamma + C(\Lambda)(1+B^\circ(x_0-x))^{-1} F(x,t,c_0(\Lambda)B^\circ(x_0-x)\nabla B^\circ(x_0-x), -I) \\ &\quad \leq -\Gamma + C(\Lambda) \leq 0 \end{aligned}$$

when $\Gamma \geq C(\Lambda)$. Here we have used

$$F(x,t,p,-I) \leq C(\Lambda)(1+|p|), \tag{50}$$

which can easily be derived from the definition of F and (8).

To prove that v is a subsolution of

$$\partial_t v + F_\star(.,.,\nabla v, D^2 v) \leq 0 \tag{51}$$

on the whole $\mathbb{R}^n \times]0,T[$, we consider $\psi \in C^{2,1}(\mathbb{R}^n \times\,]0,T[)$ with

$$(v-\psi) \leq (v-\psi)(x_0,s_0)$$

for some $0 < s_0 < T$. We get from (49) that

$$\nabla\psi(x_0,s_0) = \nabla v(x_0,s_0) = 0.$$

Adding $|x-x_0|^4 + |t-s_0|^2$ to ψ, we may assume $(v-\psi)(x,t) < (v-\psi)(x_0,s_0)$ for all $(x,t) \neq (x_0,s_0)$. We define $\psi_\tau(x,t) := \psi(x,t) + \tau\ x * N$ for some $N \neq 0$. As $\tau \to 0$, we get

$$(v-\psi_\tau)(.,s_0) \leq (v-\psi_\tau)(x_\tau,s_0)$$

on a neighborhood $U(x_0)$ of x_0 and $x_\tau \to x_0$. This yields

$$\nabla v(x_\tau,s_0) = \nabla\psi_\tau(x_\tau,s_0) = \nabla\psi(x_\tau,s_0) + \tau N \neq \nabla\psi(x_\tau,s_0);$$

hence $x_\tau \neq x_0$. Furthermore, we have

$$D^2\psi(x_0,s_0) \leftarrow D^2\psi_\tau(x_\tau,s_0) \geq D^2 v(x_\tau,s_0) \geq -(1+B^\circ(x_0-x_\tau))^{-1}C(\Lambda)I \to -C(\Lambda)I.$$

Using (50), we get

$$\begin{aligned} &\partial_t\psi(x_0,s_0) + F_\star(x_0,s_0,\nabla\psi(x_0,s_0), D^2\psi(x_0,s_0)) \\ &\quad \leq -\Gamma + F_\star(x_0,s_0,0,-C(\Lambda)I) \leq -\Gamma + C(\Lambda) \leq 0; \end{aligned}$$

hence (51) is established.

We define $U := \{x \in \mathbb{R}^n \mid B^\circ(x_0-x) < \varrho\}$. For $x \in U$, we see that

$$d(x_0,x) = B^\circ(x_0-x) < \varrho = \delta(x_0,t_0);$$

hence

$$\omega(x, t_0) > 0$$

and

$$v(x, t_0) \leq \mu(\varrho) \leq \mu(\varrho)\omega_\infty(x, t_0).$$

For $x \in \partial U$ and $t \in [t_0, T[$, we get $v(x, t) \leq 0 \leq \mu(\varrho)\omega_\infty(x, t)$. We conclude with the comparison principle (Theorem 2.4),

$$\forall x \in U: \quad \forall t \in [t_0, T[: \quad v(x, t) \leq \mu(\varrho)\omega_\infty(x, t). \tag{52}$$

We choose $y_0 \in [\omega(., t_1) \leq 0]$ such that $\delta(x_0, t_1) = B^\circ(x_0 - y_0)$. If $y_0 \notin U$, we see that $\delta(x_0, t_1) = B^\circ(x_0 - y_0) \geq \varrho = \delta(x_0, t_0)$ and

$$\mu(\delta(x_0, t_1)) \geq \mu(\varrho) \geq \mu(\delta(x_0, t_0)).$$

If $y_0 \in U$, we obtain from (52)

$$0 = \mu(\varrho)\omega_\infty(y_0, t_1) \geq v(y_0, t_1) = \mu(\varrho) - \Gamma(t_1 - t_0) - \mu(B^\circ(x_0 - y_0)).$$

Taking into account that $\varrho = \delta(x_0, t_0)$ and $B^\circ(x_0 - y_0) = \delta(x_0, t_1)$, (47) follows. □

Proposition 3.5. *δ is continuous from below; that is, if $(x_j, t_j) \to (x_0, t_0)$ and $t_j \leq t_0 < T$, then*

$$\delta(x_j, t_j) \to \delta(x_0, t_0).$$

Proof. Because of the already established lower semicontinuity of δ, it suffices to show

$$\limsup_{j \to \infty} \delta(x_j, t_j) \leq \delta(x_0, t_0). \tag{53}$$

From (47), we get

$$\mu(\delta(x_0, t_0)) \geq \limsup_{j \to \infty}(\mu(\delta(x_0, t_j)) - C(\Lambda)(t_0 - t_j)) = \mu\left(\limsup_{j \to \infty} \delta(x_0, t_j)\right),$$

where μ is defined in (3.4), since μ is continuous and increasing. Since μ is strictly increasing, it follows that

$$\delta(x_0, t_0) \geq \limsup_{j \to \infty} \delta(x_0, t_j),$$

which yields (53) when taking into account that $|\delta(x_j, t_j) - \delta(x_0, t_j)| \leq C(\Lambda)|x_j - x_0| \to 0$. □

Proposition 3.6. *δ is a viscosity supersolution of*

$$\begin{aligned} B(\nabla\delta) &\geq 1, & -B(\nabla\delta) &\geq -1, \\ -\nabla B(\nabla\delta) D^2\delta \nabla B(\nabla\delta) &\geq 0, & -D^2\delta &\geq -C(\Lambda)\delta^{-1} I, \end{aligned}$$

and

$$\partial_t \delta \geq -C(\Lambda)\left(1 + \frac{1}{\delta}\right)$$

in $[\delta > 0] \cap (\mathbb{R}^n \times]0, T[)$.

Proof. Let $\varphi \in C^{2,1}(\mathbb{R}^n \times]0,T[)$, $(x_0, t_0) \in \mathbb{R}^n \times]0,T[$ with $\delta(x_0,t_0) > 0$ and $(\delta - \varphi) \geq (\delta - \varphi)(x_0, t_0) = 0$ in $\mathbb{R}^n \times]0,T[$. We choose $y_0 \in [\omega(.,t_0) \leq 0]$ with

$$0 < \delta(x_0,t_0) = d(x_0,y_0) = B^\circ(x_0 - y_0);$$

note in particular that $x_0 \neq y_0$. For all $x \in \mathbb{R}^n$, we get $B^\circ(x - y_0) \geq \delta(x, t_0)$; hence

$$B^\circ(x - y_0) - \varphi(x,t_0) \geq B^\circ(x_0 - y_0) - \varphi(x_0,t_0).$$

Since $x_0 \neq y_0$ and $B^\circ \in C^2(\mathbb{R}^n - \{0\})$, we get

$$\nabla\varphi(x_0,t_0) = \nabla B^\circ(x_0 - y_0)$$

and

$$D^2\varphi(x_0,t_0) \leq D^2 B^\circ(x_0 - y_0) \leq C(\Lambda) I |x_0 - y_0|^{-1} = C(\Lambda)\delta(x_0,t_0)^{-1} I,$$

where we have used the fact that $D^2 B^\circ$ is homogeneous of degree -1. From (39), we get

$$B(\nabla\varphi(x_0,t_0)) = 1$$

and

$$-\nabla B(\nabla\varphi) D^2\varphi \nabla B(\nabla\varphi)(x_0,t_0) \geq 0.$$

From (3.4), for $0 \leq t < t_0$, we get

$$\begin{aligned}\mu(\varphi(x_0,t_0)) = \mu(\delta(x_0,t_0)) &\geq \mu(\delta(x_0,t)) - C(\Lambda)(t_0 - t)\\ &\geq \mu(\varphi(x_0,t)) - C(\Lambda)(t_0 - t),\end{aligned}$$

and

$$\begin{aligned}-C(\Lambda) \leq \frac{\mu(\varphi(x_0,t)) - \mu(\varphi(x_0,t_0))}{t - t_0} &\to \mu'(\varphi(x_0,t_0))\partial_t\varphi(x_0,t_0)\\ = \frac{\varphi(x_0,t_0)}{1 + \varphi(x_0,t_0)}\partial_t\varphi(x_0,t_0).&\end{aligned}$$

Since $\varphi(x_0,t_0) = \delta(x_0,t_0)$, this yields

$$\partial_t\varphi(x_0,t_0) \geq -C(\Lambda)(1 + \delta(x_0,t_0)^{-1}). \qquad \square$$

4. Convergence.

Remark 4.1. In this section, we construct sub- and supersolutions for the double-obstacle Allen–Cahn problem which satisfy the additional condition of the modified comparison principle (Theorem 2.12).

We consider β, A, B, and u as in (8) and ω and δ as in section 3.2.

Definition 4.2. *We define the following auxiliary functions, which appear in formal asymptotics for the double-obstacle Allen–Cahn equation:*

$$\psi_0(r) := \begin{cases} 1, & r \geq \frac{\pi}{2},\\ \sin(r), & |r| \leq \frac{\pi}{2},\\ -1, & r \leq -\frac{\pi}{2}\end{cases}$$

and

$$\psi_1(r) := \begin{cases} \dfrac{1}{2}(r\psi_0(r) - \dfrac{\pi}{2} + \psi_0'(r)), & |r| \leq \dfrac{\pi}{2}, \\ 0, & |r| \geq \dfrac{\pi}{2}. \end{cases}$$

We see that $\psi_0,\ \psi_1 \in \mathrm{C}^2([-\frac{\pi}{2}, \frac{\pi}{2}]) \cap \mathrm{C}^{1,1}(\mathbb{R})$. *Moreover,*

$$\begin{aligned} &\psi_0'' + \psi_0 = 0 \quad \textit{on} \ \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \quad \textit{and} \\ &\psi_1'' + \psi_1 = \psi_0' - \frac{\pi}{4} \quad \textit{on} \ \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \end{aligned} \tag{54}$$

We define

$$\psi(r, v) := \psi_0(r) + \varepsilon v \psi_1(r)$$

and

$$\lambda_\varepsilon(x, t) := \frac{\delta(x, t)}{\varepsilon} - \pi - f(t),$$

where $f(t) := \alpha \exp(-\gamma^2 t)$ *with* $1 \leq \alpha,\ \gamma \leq C(\Lambda)$ *chosen below. We set* $v := u + \varepsilon g$, *where* $g(t) := \alpha\gamma \exp(-\gamma^2 t)$, *and*

$$\psi_\varepsilon := \psi(\lambda_\varepsilon, v).$$

PROPOSITION 4.3. *For* $0 < \varepsilon < \varepsilon_0(\Lambda)$, ψ_ε *is a supersolution of*

$$\max\left(\psi_\varepsilon - 1, \min\left(\psi_\varepsilon + 1, \partial_t \psi_\varepsilon + \frac{1}{\varepsilon} G_\varepsilon^\star(.,., \psi_\varepsilon, \nabla\psi_\varepsilon, D^2\psi_\varepsilon)\right)\right) \geq 0 \quad \textit{in } \mathbb{R}^n \times]0, T[. \tag{55}$$

Moreover, ψ_ε *satisfies the additional condition* (26) *of the modified comparison principle*; *that is, for* $(a, 0, X) \in \overline{\mathrm{P}}^{2,-}\psi_\varepsilon(x_0, t_0)$, $0 < t_0 < T$, *and* $\psi_\varepsilon(x_0, t_0) < 1$,

$$\limsup_{p \to 0,\ p \neq 0} \left(\varepsilon\beta^* a - \varepsilon tr(D^2 A(p) X) - \frac{1}{\varepsilon}\psi_\varepsilon(x_0, t_0) - \frac{\pi}{4} u(x_0, t_0)\right) \geq 0 \tag{56}$$

holds for any $\inf \beta \leq \beta^\star \leq \sup \beta$.

Proof. We take ε so small that $\varepsilon f,\ \varepsilon g,\ \varepsilon|f'|,\ \varepsilon|g'| \leq 1$. We have $|\psi_1(r)|,\ |\psi_1'(r)| \leq \psi_0'(r)$. For $|v| \leq C(\Lambda)$, this yields

$$\psi_r(r, v) = \psi_0'(r) + \varepsilon v \psi_1'(r) = \psi_0'(r)(1 + O_\Lambda(\varepsilon));$$

hence

$$\frac{1}{2}\psi_0'(r) \leq \psi_r(r, v) \leq 2\psi_0'(r) \tag{57}$$

when $0 < \varepsilon < \varepsilon_0(\Lambda)$. Therefore, $\psi_\varepsilon \in \mathrm{LSC}(\mathbb{R}^n \times [0, T[)$ since $\delta \in \mathrm{LSC}(\mathbb{R}^n \times [0, T[)$. Since $\psi(r, v) = -1$ for $r \leq -\frac{\pi}{2}$, we get

$$\psi_\varepsilon \geq -1.$$

Suppose $(\bar{x},\bar{t}) \in \mathbb{R}^n \times]0,T[$ is such that $\psi_\varepsilon(\bar{x},\bar{t}) < 1$. It follows that $\lambda_\varepsilon(\bar{x},\bar{t}) < \frac{\pi}{2}$. From Remark 2.3, we see that for any $(\bar{a},\bar{p},\bar{X}) \in \mathrm{P}^{2,-}\psi_\varepsilon(\bar{x},\bar{t})$, there is a $\varphi \in C^{2,1}(\mathbb{R}^n \times [0,T[)$ such that

$$(58) \qquad \begin{array}{l} (\bar{a},\bar{p},\bar{X}) = (\partial_t\varphi, \nabla\varphi, D^2\varphi)(\bar{x},\bar{t}) \quad \text{and} \\ \psi_\varepsilon - \varphi \geq (\psi_\varepsilon - \varphi)(\bar{x},\bar{t}), \end{array}$$

and, without loss of generality, $(\psi_\varepsilon - \varphi)(\bar{x},\bar{t}) = 0$ and $(\bar{x},\bar{t})$ is a strict minimum. Our aim is to prove that ψ_ε is a supersolution. By Definition 2.1, it remains to show that

$$(59) \qquad \bar{a} + \frac{1}{\varepsilon} G_\varepsilon^\star(\bar{x},\bar{t},\psi_\varepsilon(\bar{x},\bar{t}),\bar{p},\bar{X})) \geq 0.$$

We distinguish two cases.

(i) $\lambda_\varepsilon(\bar{x},\bar{t}) < -\frac{\pi}{2}$. Using the continuity from below in time of δ (see Proposition 3.5), we conclude that

$$\lambda_\varepsilon(x,t) < -\frac{\pi}{2} \quad \text{for } (x,t) \in U(\bar{x},\bar{t}) \text{ and } t \leq \bar{t},$$

and for these (x,t), it follows that $\psi_\varepsilon(x,t) = -1$. This yields

$$\bar{X} = D^2\varphi(\bar{x},\bar{t}) \leq 0, \qquad \bar{a} = \partial_t\varphi(\bar{x},\bar{t}) \geq 0,$$

and

$$\bar{p} = \nabla\varphi(\bar{x},\bar{t}) = 0.$$

For $p \neq 0$ and $\inf\beta \leq \beta^\star \leq \sup\beta$, we obtain

$$(60) \qquad \varepsilon\beta^\star\bar{a} - \varepsilon\mathrm{tr}(D^2A(p)\bar{X}) - \frac{1}{\varepsilon}\psi_\varepsilon(\bar{x},\bar{t}) - \frac{\pi}{4}u(\bar{x},\bar{t}) \geq \left(\frac{1}{\varepsilon} - C(\Lambda)\right) \geq 0$$

when $0 < \varepsilon < \varepsilon_0(\Lambda)$. Taking $\beta^\star = \beta(p)$ and letting p tend to $\bar{p} = 0$, we obtain

$$\begin{aligned} &\bar{a} + \frac{1}{\varepsilon}G_\varepsilon^\star(\bar{x},\bar{t},\psi_\varepsilon(\bar{x},\bar{t}),\bar{p},\bar{X}) \\ &\quad \geq \limsup_{p\to 0, p\neq 0}\left(\bar{a} + \frac{1}{\varepsilon}G_\varepsilon(\bar{x},\bar{t},\psi_\varepsilon(\bar{x},\bar{t}),p,\bar{X})\right) \\ &\quad = \limsup_{p\to 0, p\neq 0}\left(\bar{a} - \frac{1}{\varepsilon\beta(p)}\left(\varepsilon\mathrm{tr}(D^2A(p)\bar{X}) + \frac{1}{\varepsilon}\psi_\varepsilon(\bar{x},\bar{t}) + \frac{\pi}{4}u(\bar{x},\bar{t})\right)\right) \geq 0, \end{aligned}$$

which is (59).

(ii) $-\frac{\pi}{2} \leq \lambda_\varepsilon(\bar{x},\bar{t}) < \frac{\pi}{2}$. In the next two subsections, we will establish the existence of subsequences (x_τ,t_τ) and $\psi_\varepsilon^\tau(x_\tau,t_\tau)$ such that as $\tau \to 0$,

$$(61) \qquad \begin{array}{l} (x_\tau,t_\tau) \to (\bar{x},\bar{t}), \\ \psi_\varepsilon^\tau(x_\tau,t_\tau) \to \psi_\varepsilon(\bar{x},\bar{t}), \\ \nabla\varphi(x_\tau,t_\tau) \neq 0, \end{array}$$

and

$$R_\varepsilon^\tau := \left(\varepsilon\beta^\star\partial_t\varphi - \varepsilon\mathrm{tr}(D^2A(\nabla\varphi)D^2\varphi) - \frac{1}{\varepsilon}\psi_\varepsilon^\tau - \frac{\pi}{4}u)(x_\tau,t_\tau\right) \geq \varepsilon - \varepsilon^{-1}\varrho_\Lambda(\tau),$$

where $\inf\beta \le \beta^\star \le \sup\beta$ and $\beta^\star = \beta(\nabla\varphi(x_\tau, t_\tau))$. Furthermore, if $|\nabla\varphi(x_\tau, t_\tau)| \le \Lambda\varepsilon$, then the last inequality holds for any $\inf\beta \le \beta^\star \le \sup\beta$. Here $\varrho_\Lambda(\tau) \to 0$.

Taking $\beta^\star = \beta(\nabla\varphi(x_\tau, t_\tau))$ in the definition of R^τ_ε, we find from (61) that

$$\begin{aligned}
&\varepsilon\partial_t\varphi(\bar{x},\bar{t}) + G^\star_\varepsilon(\bar{x},\bar{t},\psi_\varepsilon(\bar{x},\bar{t}),\nabla\varphi(\bar{x},\bar{t}),D^2\varphi(\bar{x},\bar{t}))\\
&\quad \ge \limsup_{\tau\to0}(\varepsilon\partial_t\varphi(x_\tau,t_\tau) + G_\varepsilon(x_\tau,t_\tau,\psi^\tau_\varepsilon(x_\tau,t_\tau),\nabla\varphi(x_\tau,t_\tau),D^2\varphi(x_\tau,t_\tau)))\\
&\quad = \limsup_{\tau\to0}(\beta(\nabla\varphi(x_\tau,t_\tau))^{-1}R^\tau_\varepsilon) \ge 0,
\end{aligned}$$

which is (59). Thus ψ_ε is a supersolution.

We now turn to the proof of (56). First, we observe that for any $(\bar{x},\bar{t}) \in \mathbb{R}^n \times]0,T[$ with $\psi_\varepsilon(\bar{x},\bar{t}) < 1$ and φ satisfying (58) with $|\bar{p}| = |\nabla\varphi(\bar{x},\bar{t})| \le \Lambda\varepsilon$,

$$\limsup_{q\to\bar{p},q\neq0}\left(\varepsilon\beta^\star\bar{a} - \varepsilon\operatorname{tr}(D^2A(q)\bar{X}) - \frac{1}{\varepsilon}\psi_\varepsilon(\bar{x},\bar{t}) - \frac{\pi}{4}u(\bar{x},\bar{t})\right) \ge 0 \tag{62}$$

for $\inf\beta \le \beta^\star \le \sup\beta$.

To prove (62), we again distinguish two cases. In case (i), $\lambda_\varepsilon(\bar{x},\bar{t}) < -\frac{\pi}{2}$, (62) is an immediate consequence of (60). In case (ii), $-\frac{\pi}{2} \le \lambda_\varepsilon(\bar{x},\bar{t}) < -\frac{\pi}{2}$, it follows from (61), for any $\inf\beta \le \beta^\star \le \sup\beta$, that

$$\begin{aligned}
&\limsup_{q\to\bar{p},q\neq0}\left(\varepsilon\beta^\star\bar{a} - \varepsilon\operatorname{tr}(D^2A(q)\bar{X}) - \frac{1}{\varepsilon}\psi_\varepsilon(\bar{x},\bar{t}) - \frac{\pi}{4}u(\bar{x},\bar{t})\right)\\
&\quad\ge \limsup_{q\to\nabla\varphi(\bar{x},\bar{t}),q\neq0}\left(\varepsilon\beta^\star\partial_t\varphi(\bar{x},\bar{t}) - \varepsilon\operatorname{tr}(D^2A(q)D^2\varphi(\bar{x},\bar{t})) - \frac{1}{\varepsilon}\psi_\varepsilon(\bar{x},\bar{t}) - \frac{\pi}{4}u(\bar{x},\bar{t})\right)\\
&\quad\ge \limsup_{\tau\to0} R^\tau_\varepsilon \ge 0,
\end{aligned}$$

which is (62).

Now we consider $(a,0,X) \in \overline{\mathrm{P}}^{2,-}\psi_\varepsilon(x_0,t_0)$ with $\psi_\varepsilon(x_0,t_0) < 1$. From the definition of $\overline{\mathrm{P}}^{2,-}$, we get $(a_j,p_j,X_j) \in \mathrm{P}^{2,-}\psi_\varepsilon(x_j,t_j)$, which, as $j\to\infty$, satisfies

$$\begin{aligned}
(a_j,p_j,X_j) &\to (a,0,X),\\
(x_j,t_j) &\to (x_0,t_0),
\end{aligned}$$

and

$$\psi_\varepsilon(x_j,t_j) \to \psi_\varepsilon(x_0,t_0).$$

We apply (62) to $(\bar{x},\bar{t}) = (x_j,t_j)$ and obtain

$$\limsup_{q\to p_j,\ q\neq0}\left(\varepsilon\beta^\star a_j - \varepsilon\operatorname{tr}(D^2A(q)X_j) - \frac{1}{\varepsilon}\psi_\varepsilon(x_j,t_j) - \frac{\pi}{4}u(x_j,t_j)\right) \ge 0,$$

which yields the existence of $q_j \neq 0$ with $q_j \to 0$ and

$$\varepsilon\beta^\star a_j - \varepsilon\operatorname{tr}(D^2A(q_j)X_j) - \frac{1}{\varepsilon}\psi_\varepsilon(x_j,t_j) - \frac{\pi}{4}u(x_j,t_j) \ge -\frac{1}{j}.$$

From this we infer (56), concluding the proof. □

4.1. Approximation. In this section, we approximate ψ_ε by a smooth ψ_ε^τ, and we will get $(x_\tau, t_\tau) \to (\bar{x}, \bar{t})$ as in (61).

We take the notation of the preceding subsection, and we assume $-\frac{\pi}{2} \leq \lambda_\varepsilon(\bar{x}, \bar{t}) < \frac{\pi}{2}$.

From the definition of λ_ε, it follows that $\frac{\pi}{2}\varepsilon < \delta(\bar{x}, \bar{t})$, and from the lower semicontinuity of δ (see Proposition 3.3), we get

$$\begin{aligned} \delta &\geq \frac{\pi}{2}\varepsilon, \\ \lambda_\varepsilon &\geq -\pi \end{aligned} \quad \text{in } \overline{U(\bar{x}, \bar{t})} \tag{63}$$

for some neighborhood $U(\bar{x}, \bar{t})$ of $(\bar{x}, \bar{t})$. We take a Dirac sequence $\eta_\tau(r) = \tau^{-1}\eta(\frac{r}{\tau})$ for $\eta \in \mathrm{C}_0^\infty(\mathbb{R}),\ \eta \geq 0, \int \eta = 1$, and define

$$\begin{aligned} \psi_0^\tau(r) &:= \textstyle\int \psi_0(s)\eta_\tau(r-s)ds, \\ \psi_1^\tau(r) &:= \textstyle\int \psi_1(s)\eta_\tau(r-s)ds. \end{aligned}$$

We get $(\psi_0^\tau)',\ \psi_1^\tau \in \mathrm{C}_0^\infty(\mathbb{R})$, and

$$|\psi_1^\tau|,\ |(\psi_1^\tau)'| \leq C(\Lambda)(\psi_0^\tau)'. \tag{64}$$

We define

$$\psi^\tau(r, v) := \psi_0^\tau(r) + \varepsilon v \psi_1^\tau(r) + \tau r,$$

and for $|v| \leq C(\Lambda),\ 0 < \varepsilon < \varepsilon_0(\Lambda)$, we get

$$\psi_r^\tau(r, v) \geq \tau > 0. \tag{65}$$

We choose $u^\tau \in C^\infty(\mathbb{R}^n \times [0, T])$ with

$$\|u^\tau\|_{C^{2,1}(\mathbb{R} \times [0,T])} \leq C(\Lambda)$$

and

$$\|u^\tau - u\|_{L^\infty(\mathbb{R}^n \times [0,T])} \leq \varrho(\tau) \to 0 \quad \text{for } \tau \to 0.$$

We define $v^\tau := u^\tau + \varepsilon g$ and get

$$\|v^\tau\|_{C^{2,1}(\mathbb{R}^n \times [0,T])} \leq C(\Lambda).$$

We set

$$\psi_\varepsilon^\tau := \psi^\tau(\lambda_\varepsilon, v^\tau).$$

Using the above approximation properties in τ and (63), we obtain $\varrho_\Lambda(\tau) \to 0$ as $\tau \to 0$ such that

$$\begin{aligned} \psi_\varepsilon^\tau - \psi_\varepsilon \geq &-\|\psi_0^\tau - \psi_0\|_{L^\infty(\mathbb{R})} - \varepsilon\|v^\tau - v\|_{L^\infty(\mathbb{R}^n \times [0,T])}\|\psi_1^\tau\|_{L^\infty(\mathbb{R})} \\ &- \varepsilon\|v\|_{L^\infty(\mathbb{R}^n \times [0,T])}\|\psi_1^\tau - \psi_1\|_{L^\infty(\mathbb{R})} - \tau\pi \\ \geq &-\varrho_\Lambda(\tau) \end{aligned}$$

on $\overline{U(\bar{x}, \bar{t})}$ and

$$|\psi_\varepsilon^\tau(\bar{x}, \bar{t}) - \psi_\varepsilon(\bar{x}, \bar{t})| \leq \varrho_\Lambda(\tau).$$

Therefore, for small τ,

$$\begin{gathered}\exists (x_\tau, t_\tau) \in U(\bar{x}, \bar{t})\colon \quad \psi_\varepsilon^\tau - \varphi \geq (\psi_\varepsilon^\tau - \varphi)(x_\tau, t_\tau) \quad \text{on } U(\bar{x}, \bar{t}), \\ (x_\tau, t_\tau) \to (\bar{x}, \bar{t}) \quad \text{as } \tau \to 0, \quad \text{and} \\ (\psi_\varepsilon^\tau - \varphi)(x_\tau, t_\tau) =: \nu_\tau \to (\psi_\varepsilon - \varphi)(\bar{x}, \bar{t}) = 0.\end{gathered} \tag{66}$$

The last convergence yields

$$\begin{aligned}1 > \psi_\varepsilon(\bar{x}, \bar{t}) &= \lim_{\tau \to 0} \psi_\varepsilon^\tau(x_\tau, t_\tau) \\ &\geq \limsup_{\tau \to 0} \psi(\lambda_\varepsilon(x_\tau, t_\tau), v(x_\tau, t_\tau)) \geq \psi\left(\limsup_{\tau \to 0} \lambda_\varepsilon(x_\tau, t_\tau), v(\bar{x}, \bar{t})\right);\end{aligned}$$

hence

$$\lambda_\varepsilon(x_\tau, t_\tau) < \frac{\pi}{2} \quad \text{for } \tau \text{ small.} \tag{67}$$

Because of (65) and since ψ^τ is smooth, there is $\delta^\tau \in C^{2,1}(U(\bar{x}, \bar{t}))$ such that

$$\begin{gathered}\varphi(x, t) = \psi^\tau(\lambda_\varepsilon^\tau(x, t), v^\tau(x, t)) - \nu_\tau \quad \text{and} \\ \lambda_\varepsilon^\tau(x, t) = \frac{\delta^\tau(x, t)}{\varepsilon} - \pi - f(t),\end{gathered} \tag{68}$$

where ν_τ is defined in (66). For $(x, t) \in U(\bar{x}, \bar{t})$, we obtain

$$\psi^\tau(\lambda_\varepsilon, v^\tau) - \varphi(x, t) = (\psi_\varepsilon^\tau - \varphi)(x, t) \geq \nu_\tau = (\psi^\tau(\lambda_\varepsilon^\tau, v^\tau) - \varphi)(x, t),$$

and equality holds for $(x, t) = (x_\tau, t_\tau)$. From (65), we conclude

$$\begin{gathered}\delta \geq \delta^\tau \quad \text{in } U(\bar{x}, \bar{t}) \quad \text{and} \\ \delta(x_\tau, t_\tau) = \delta^\tau(x_\tau, t_\tau) \geq \frac{\pi}{2}\varepsilon,\end{gathered} \tag{69}$$

where we have used (63).

4.2. Computation. In this section, we will carry out the computations to establish the fourth line of (61). We again use the notation of the preceding sections.

We continue from (69). Observing Remark 2.3, we have that

$$(\partial_t \delta^\tau, \nabla \delta^\tau, D^2 \delta^\tau)(x_\tau, t_\tau) \in \mathrm{P}^{2,-}\delta(x_\tau, t_\tau),$$

and from the definition of supersolutions, using Proposition 3.6, we obtain

$$\begin{gathered}B(\nabla \delta^\tau(x_\tau, t_\tau)) = 1, \qquad c_0(\Lambda) \leq |\nabla \delta^\tau(x_\tau, t_\tau)| \leq C(\Lambda), \\ -\nabla B(\nabla \delta^\tau) D^2 \delta^\tau \nabla B(\nabla \delta^\tau)(x_\tau, t_\tau) \geq 0, \\ D^2 \delta^\tau(x_\tau, t_\tau) \leq C(\Lambda) \delta(x_\tau, t_\tau)^{-1} I \leq C(\Lambda) \varepsilon^{-1} I, \quad \text{and} \\ \partial_t \delta^\tau(x_\tau, t_\tau) \geq -C(\Lambda)(1 + \delta(x_\tau, t_\tau)^{-1}) \geq -C(\Lambda) \varepsilon^{-1}.\end{gathered} \tag{70}$$

Further, from Proposition 3.3, we get

$$(\beta(\nabla \delta^\tau) \partial_t \delta^\tau - \operatorname{tr}(B(\nabla \delta^\tau) D^2 B(\nabla \delta^\tau) D^2 \delta^\tau) - B(\nabla \delta^\tau) u + C(\Lambda) |\nabla \delta^\tau| \delta)(x_\tau, t_\tau) \geq 0,$$

and, since $D^2 A = B D^2 B + B' \otimes B'$, we conclude

$$(\beta(\nabla \delta^\tau) \partial_t \delta^\tau - \operatorname{tr}(D^2 A(\nabla \delta^\tau) D^2 \delta^\tau) - u + C(\Lambda) \delta)(x_\tau, t_\tau) \geq 0. \tag{71}$$

From (68), we compute

$$\nabla\varphi = \varepsilon^{-1}\psi_r^\tau \nabla\delta^\tau + \varepsilon\psi_1^\tau \nabla v^\tau = \varepsilon^{-1}\psi_r^\tau(\nabla\delta^\tau + O_\Lambda(\varepsilon^2)). \tag{72}$$

From (65) and $|\nabla\delta^\tau(x_\tau, t_\tau)| \geq c_0(\Lambda)$, we get

$$\nabla\varphi(x_\tau, t_\tau) \neq 0, \tag{73}$$

which is the third line of (61).

Differentiating again, we get

$$\begin{aligned} D^2\varphi = {}& \varepsilon^{-2}\psi_{rr}^\tau \nabla\delta^\tau \otimes \nabla\delta^\tau + \varepsilon^{-1}\psi_r^\tau D^2\delta^\tau \\ &+ (\psi_1^\tau)'(\nabla v^\tau \otimes \nabla\delta^\tau + \nabla\delta^\tau \otimes \nabla v^\tau) + \varepsilon\psi_1^\tau D^2 v^\tau \\ = {}& \varepsilon^{-2}\psi_{rr}^\tau \nabla\delta^\tau \otimes \nabla\delta^\tau + \varepsilon^{-1}\psi_r^\tau D^2\delta^\tau + O_\Lambda(1) \end{aligned} \tag{74}$$

since $|\nabla\delta^\tau(x_\tau, t_\tau)| \leq C(\Lambda)$.

Further, we obtain

$$\varepsilon\partial_t\varphi = \psi_r^\tau(\partial_t\delta^\tau - \varepsilon f') + \varepsilon^2\psi_1^\tau\partial_t v^\tau = \psi_r^\tau(\partial_t\delta^\tau - \varepsilon f' + O_\Lambda(\varepsilon^2)). \tag{75}$$

We recall the definition of R_ε^τ in (61),

$$R_\varepsilon^\tau := \left(\varepsilon\beta^\star\partial_t\varphi - \varepsilon\operatorname{tr}(D^2A(\nabla\varphi)D^2\varphi) - \frac{1}{\varepsilon}\psi_\varepsilon^\tau - \frac{\pi}{4}u\right)(x_\tau, t_\tau) \tag{76}$$

for $\inf\beta \leq \beta^\star \leq \sup\beta$ and $\beta^\star = \beta(\nabla\varphi(x_\tau, t_\tau))$ or $|\nabla\varphi(x_\tau, t_\tau)| \leq \Lambda\varepsilon$.

Using (72), (74), and (75), we get

$$\begin{aligned} R_\varepsilon^\tau \geq {}& -\varepsilon^{-1}(\psi_{rr}^\tau \nabla\delta^\tau D^2A(\nabla\varphi)\nabla\delta^\tau + \psi^\tau) \\ &+ \psi_r^\tau\beta^\star\partial_t\delta^\tau - \frac{\pi}{4}u \\ &- \psi_r^\tau\operatorname{tr}(D^2A(\nabla\varphi)D^2\delta^\tau) \\ &- \varepsilon\psi_r^\tau\beta^\star f' - C(\Lambda)\varepsilon. \end{aligned} \tag{77}$$

Using the homogeneity of A and β, (8), and (72), we obtain

$$\begin{aligned} D^2A(\nabla\varphi) &= D^2A(\nabla\delta^\tau + O_\Lambda(\varepsilon^2)) = D^2A(\nabla\delta^\tau) + O_\Lambda(\varepsilon^2), \\ \nabla\delta^\tau D^2A(\nabla\delta^\tau)\nabla\delta^\tau &= 2A(\nabla\delta^\tau) = B(\nabla\delta^\tau)^2 = 1, \quad \text{and} \\ \beta(\nabla\varphi) &= \beta(\nabla\delta^\tau + O_\Lambda(\varepsilon^2)) = \beta(\nabla\delta^\tau) + O_\Lambda(\varepsilon^2). \end{aligned} \tag{78}$$

From (77) and (78), it follows that

$$\begin{aligned} R_\varepsilon^\tau \geq {}& -\varepsilon^{-1}(\psi_{rr}^\tau + \psi^\tau) \\ &+ \psi_r^\tau(\beta^\star\partial_t\delta^\tau - \operatorname{tr}(D^2A(\nabla\varphi)D^2\delta^\tau) - u) \\ &+ \left(\psi_r^\tau - \frac{\pi}{4}\right)u \\ &- \varepsilon c_0(\Lambda)\psi_r^\tau f' - C(\Lambda)\varepsilon, \end{aligned} \tag{79}$$

where again $|\nabla\delta^\tau(x_\tau, t_\tau)| \leq C(\Lambda)$ was used.

In the following lemma, we will prove that

$$\psi_r^\tau(\beta^\star\partial_t\delta^\tau - \operatorname{tr}(D^2A(\nabla\varphi)D^2\delta^\tau) - u) \geq -C(\Lambda)(\delta + \varepsilon). \tag{80}$$

From (63) and (67), we know that $|\lambda_\varepsilon(x_\tau, t_\tau)| \leq \pi$. This yields $\delta(x_\tau, t_\tau) \leq \varepsilon(\lambda_\varepsilon(x_\tau, t_\tau) + \pi + f(t_\tau)) \leq C(\Lambda)\varepsilon(1 + f(t_\tau))$. Plugging (80) in (79), we obtain

$$
\begin{aligned}
R_\varepsilon^\tau \geq & - \varepsilon^{-1}(\psi_{rr}^\tau + \psi^\tau) \\
& + \left(\psi_r^\tau - \frac{\pi}{4}\right) u \\
& - \varepsilon \psi_r^\tau c_0(\Lambda) f' - C(\Lambda)\varepsilon(1+f) \\
\geq & -\varepsilon^{-1}(\psi_{rr}^\tau + \psi^\tau) \\
& + \left((\psi_0^\tau)' - \frac{\pi}{4}\right) u \\
& - \varepsilon(\psi_0^\tau)' c_0(\Lambda) f' - C(\Lambda)\varepsilon(1+f)
\end{aligned} \tag{81}
$$

since $|\varepsilon f'|, |\varepsilon f| \leq 1$.

Setting $r := \lambda_\varepsilon(x_\tau, t_\tau)$, we compute

$$
\begin{aligned}
&-\varepsilon^{-1}(\psi_{rr}^\tau + \psi^\tau) + \left((\psi_0^\tau)' - \frac{\pi}{4}\right) v^\tau \\
&= \int_{\mathbb{R}} \left(-\varepsilon^{-1}(\psi_0'' + \psi_0)(s) - (\psi_1'' + \psi_1)(s) v^\tau + \left(\psi_0' - \frac{\pi}{4}\right)(s) v^\tau\right) \eta_\tau(r-s)ds - \varepsilon^{-1}\tau r \\
&= \int_{|s| \geq \frac{\pi}{2}} \left(-\varepsilon^{-1}(\psi_0'' + \psi_0)(s) - (\psi_1'' + \psi_1)(s) v^\tau + \left(\psi_0' - \frac{\pi}{4}\right)(s) v^\tau\right) \eta_\tau(r-s)ds - \varepsilon^{-1}\tau r,
\end{aligned}
$$

where we have used (54). We know from (67) and the definition of η_τ that $\eta_\tau(r-s) = 0$ when $s \geq \frac{\pi}{2}$, $r = \lambda_\varepsilon(x_\tau, t_\tau)$, and τ is small since $\limsup_{\tau \to 0} \lambda_\varepsilon(x_\tau, t_\tau) < \frac{\pi}{2}$. Therefore, the term above is estimated for small τ by

$$
\begin{aligned}
&\int_{-\infty}^{-\frac{\pi}{2}} \left(-\varepsilon^{-1}(\psi_0'' + \psi_0)(s) + \left(-\psi_1'' - \psi_1 + \psi_0 - \frac{\pi}{4}\right)(s) v^\tau\right) \eta_\tau(r-s)ds - \varepsilon^{-1}\tau r \\
&\qquad \geq \int_{-\infty}^{-\frac{\pi}{2}} (\varepsilon^{-1} - C(\Lambda)) \eta_\tau(r-s)ds - \varepsilon^{-1}\tau r \\
&\qquad \geq -\varepsilon^{-1}\varrho_\Lambda(\tau).
\end{aligned}
$$

Using the above computations, we get

$$
\begin{aligned}
R_\varepsilon^\tau \geq & \left((\psi_0^\tau)' - \frac{\pi}{4}\right)(u - v^\tau) \\
& - c_0(\Lambda)\varepsilon(\psi_0^\tau)' f' - C(\Lambda)\varepsilon(1+f) - \varepsilon^{-1}\varrho_\Lambda(\tau) \\
\geq & \, \varepsilon(\psi_0^\tau)'(-c_0(\Lambda) f' - g) \\
& + \varepsilon\left(\frac{\pi}{4} g - C(\Lambda) - C(\Lambda) f\right) - \varepsilon^{-1}\varrho_\Lambda(\tau),
\end{aligned} \tag{82}
$$

where we have used $v^\tau = u^\tau + \varepsilon g$.

Since $f(t) = \alpha \exp(-\gamma^2 t)$ and $g(t) = \alpha\gamma \exp(-\gamma^2 t)$, we obtain

$$
R_\varepsilon^\tau \geq \varepsilon - \varepsilon^{-1}\varrho_\Lambda(\tau) \tag{83}
$$

when $\gamma \geq C(\Lambda)$ and $\alpha \geq \exp(\gamma^2 T)$, which is the fourth line of (61).

Proof of (80). We now prove (80), that is,

$$
\psi_r^\tau(\beta^\star \partial_t \delta^\tau - \operatorname{tr}(D^2 A(\nabla\varphi) D^2 \delta^\tau) - u) \geq -C(\Lambda)(\delta + \varepsilon).
$$

We again take the notation of the previous subsection.

Since $\psi_r^\tau > 0$, we need only consider the situation when

$$
\beta^\star \partial_t \delta^\tau - \operatorname{tr}(D^2 A(\nabla\varphi) D^2 \delta^\tau) - u \leq 0.
$$

From (70), we know

$$\partial_t \delta^\tau \geq -C(\Lambda)\varepsilon^{-1}$$

and

$$D^2\delta^\tau \leq C(\Lambda)\varepsilon^{-1} I.$$

When $\partial_t \delta^\tau \geq 0$, since $0 \leq D^2 A \leq C(\Lambda)I$, we infer

$$c_0(\Lambda)\partial_t\delta^\tau \leq \beta^\star \partial_t \delta^\tau \leq \operatorname{tr}(D^2A(\nabla\varphi)D^2\delta^\tau) + u \leq C(\Lambda)\varepsilon^{-1};$$

hence in any case,

$$|\partial_t \delta^\tau| \leq C(\Lambda)\varepsilon^{-1}.$$

Since $D^2\delta^\tau$ is symmetric, there is an orthonormal basis $\{v_1, \dots, v_n\}$ of $\mathbb{R}^n$ and $\gamma_1, \dots, \gamma_n \in \mathbb{R}$ such that

$$D^2\delta^\tau = \sum_{i=1}^n \gamma_i v_i \otimes v_i$$

and

$$\gamma_i \leq C(\Lambda)\varepsilon^{-1}.$$

We define

$$\alpha_i^\varphi := v_i^T D^2 A(\nabla\varphi) v_i$$

and

$$\alpha_i^\delta := v_i^T D^2 A(\nabla\delta^\tau) v_i.$$

We know

$$c_0(\Lambda) \leq \alpha_i^\varphi, \qquad \alpha_i^\delta \leq C(\Lambda)$$

since $c_0(\Lambda)I \leq D^2A \leq C(\Lambda)I$. From (78), we obtain $|\alpha_i^\varphi - \alpha_i^\delta| \leq C(\Lambda)\varepsilon^2$ and

$$\exp(-C(\Lambda)\varepsilon^2) \leq \frac{\alpha_i^\varphi}{\alpha_i^\delta} \leq \exp(C(\Lambda)\varepsilon^2)$$

when $0 < \varepsilon < \varepsilon_0(\Lambda)$.

We compute

$$\begin{aligned}
&\operatorname{tr}(D^2A(\nabla\varphi)D^2\delta^\tau)\\
&\quad = \sum_{i=1}^n \gamma_i\alpha_i^\varphi = \sum_{\gamma_i>0}\gamma_i\alpha_i^\varphi + \sum_{\gamma_i<0}\gamma_i\alpha_i^\varphi\\
&\quad \leq \exp(C(\Lambda)\varepsilon^2)\sum_{\gamma_i>0}\gamma_i\alpha_i^\delta + \exp(-C(\Lambda)\varepsilon^2)\sum_{\gamma_i<0}\gamma_i\alpha_i^\delta\\
&\quad \leq \exp(-C(\Lambda)\varepsilon^2)\sum_{i=1}^n\gamma_i\alpha_i^\delta + C(\Lambda)\varepsilon = \exp(-C(\Lambda)\varepsilon^2)\operatorname{tr}(D^2A(\nabla\delta^\tau)D^2\delta^\tau) + C(\Lambda)\varepsilon.
\end{aligned}$$

Multiplying by $\exp(C(\Lambda)\varepsilon^2)$ yields

$$\begin{aligned}&\mathrm{tr}(D^2A(\nabla\varphi)D^2\delta^\tau)\\&\quad\leq \mathrm{tr}(D^2A(\nabla\delta^\tau)D^2\delta^\tau)+C(\Lambda)\varepsilon+(1-\exp(C(\Lambda)\varepsilon^2))\mathrm{tr}(D^2A(\nabla\varphi)D^2\delta^\tau).\end{aligned}$$

Taking into account the fact that $\mathrm{tr}(D^2A(\nabla\varphi)D^2\delta^\tau)\geq\beta^\star\partial_t\delta^\tau-u\geq -C(\Lambda)\varepsilon^{-1}$, we obtain

$$\mathrm{tr}(D^2A(\nabla\varphi)D^2\delta^\tau)\leq \mathrm{tr}(D^2A(\nabla\delta^\tau)D^2\delta^\tau)+C(\Lambda)\varepsilon. \tag{84}$$

In the case where $\beta^\star=\beta(\nabla\varphi)$, we observe from (78) that

$$|\beta(\nabla\varphi)\partial_t\delta^\tau-\beta(\nabla\delta^\tau)\partial_t\delta^\tau|\leq C(\Lambda)\varepsilon.$$

On the other hand, when $|\nabla\varphi|\leq\Lambda\varepsilon$, we conclude from (72) that $\psi_r^\tau\leq C(\Lambda)\varepsilon^2$; hence

$$|\psi_r^\tau(\beta^\star\partial_t\delta^\tau-\beta(\nabla\delta^\tau)\partial_t\delta^\tau)|\leq C(\Lambda)\varepsilon.$$

Together with (71) and (84), we obtain

$$\begin{aligned}&\psi_r^\tau(\beta^\star\partial_t\delta^\tau-\mathrm{tr}(D^2A(\nabla\varphi)D^2\delta^\tau)-u)\\&\quad\geq\psi_r^\tau(\beta(\nabla\delta^\tau)\partial_t\delta^\tau-\mathrm{tr}(D^2A(\nabla\delta^\tau)D^2\delta^\tau)-u)-C(\Lambda)\varepsilon\\&\quad\geq -C(\Lambda)(\varepsilon+\delta),\end{aligned}$$

which concludes the proof. □

Remark 4.4. Subsolutions can be constructed in an analogous way.

With these sub- and supersolutions and the modified comparison principle (Theorem 2.12), we are now able to prove the convergence result.

We consider the following situation. Let $D\subset\mathbb{R}^n$ be an open, periodic subset with $\emptyset\neq D$, $\overline{D}\neq\mathbb{R}^n$, and define $\Gamma_0:=\partial D$. We denote by δ_0 the signed distance function of Γ_0, positive on D, induced by the metric d of section 3.2, that is,

$$\delta_0(x):=\begin{cases}\inf\limits_{y\in\Gamma_0}d(x,y) & \text{for } x\in D,\\ -\inf\limits_{y\in\Gamma_0}d(x,y) & \text{for } x\notin D.\end{cases}$$

δ_0 is periodic, bounded, and Lipschitz continuous,

$$|\nabla\delta_0|\leq C(\Lambda).$$

Since $\emptyset\neq D$, $\overline{D}\neq\mathbb{R}^n$, for Λ large enough, we have

$$\begin{aligned}|\delta_0|&\leq\Lambda,\\ \sup\delta_0&\geq\Lambda^{-1},\quad\text{and}\\ \inf\delta_0&\leq-\Lambda^{-1}.\end{aligned}$$

From Theorem 2.5, we obtain the existence of a unique periodic $\omega\in C(\mathbb{R}^n\times[0,T[)$ which solves

$$\begin{aligned}\partial_t\omega+F(.,.,\nabla\omega,D^2\omega)&=0\quad\text{in }\mathbb{R}^n\times]0,T[\quad\text{and}\\ \omega(.,0)&=\delta_0,\end{aligned}\tag{85}$$

where F is defined in section 2.1. We assume that for $0\leq t<T$,

$$\sup\omega(.,t)\geq\Lambda^{-1}$$

and

$$\inf \omega(.,t) \leq -\Lambda^{-1}.$$

From Theorem 2.8, we get a periodic viscosity solution $\varphi_\varepsilon \in C(\mathbb{R}^n \times [0,T[)$ of the double-obstacle Allen–Cahn problem,

(86)

$$\max\left(\varphi_\varepsilon - 1, \min\left(\varphi_\varepsilon + 1, \partial_t \varphi_\varepsilon + \frac{1}{\varepsilon} G_\varepsilon(.,.,\varphi_\varepsilon, \nabla\varphi_\varepsilon, D^2\varphi_\varepsilon)\right)\right) = 0 \quad \text{in } (\mathbb{R}^n \times]0,T[).$$

For the initial conditions, we assume

$$\begin{aligned} &\varphi_\varepsilon(.,0) = 1 \quad \text{for } \delta_0 \geq C(\Lambda)\varepsilon \quad \text{and} \\ &\varphi_\varepsilon(.,0) = -1 \quad \text{for } \delta_0 \leq -C(\Lambda)\varepsilon; \end{aligned} \tag{87}$$

for example, $\varphi_\varepsilon(.,0) = \varphi_{\varepsilon,0} = \max(-1, \min(1, \frac{\delta_0}{\varepsilon}))$.

The convergence theorem can now be stated.

THEOREM 4.5.

$$\varphi_\varepsilon \to 1 \quad \textit{pointwise on } [\omega > 0]$$

and

$$\varphi_\varepsilon \to -1 \quad \textit{pointwise on } [\omega < 0].$$

Moreover, this convergence is uniform on compact subsets of $[\omega > 0]$*, respectively,* $[\omega < 0]$*.*

Proof. We define $\omega_\varepsilon^+ := \omega + \Gamma\varepsilon$ for $\Gamma = C(\Lambda)$ chosen below. According to [7], ω_ε^+ is a supersolution of (42). As in section 3.2 and Definition 4.2, we define

$$\begin{aligned} \delta_\varepsilon^+(x,t) &:= \inf_{y,\ \omega_\varepsilon^+(y,t) \leq 0} d(x,y), \\ \lambda_\varepsilon^+(x,t) &:= \frac{\delta_\varepsilon^+(x,t)}{\varepsilon} - \pi - f(t), \quad \text{and} \\ \psi_\varepsilon^+ &:= \psi(\lambda_\varepsilon^+, v). \end{aligned}$$

For $0 < \varepsilon < \varepsilon_0(\Lambda)$, we claim that

$$\psi_\varepsilon^+ \geq \varphi_\varepsilon. \tag{88}$$

From Theorem 2.12 and Proposition 4.3, it suffices to verify that

$$\psi_\varepsilon^+(.,0) \geq \varphi_\varepsilon(.,0). \tag{89}$$

When $\varphi_\varepsilon(x,0) = -1$, the inequality is satisfied since $\psi_\varepsilon^+ \geq -1$.

Now we assume $\varphi_\varepsilon(x,0) > -1$. From (87), we get $\omega(x,0) = \delta_0(x) \geq -C(\Lambda)\varepsilon$ and

$$\omega_\varepsilon^+(x,0) \geq -C(\Lambda)\varepsilon + \Gamma\varepsilon \geq \frac{\Gamma}{2}\varepsilon > 0$$

when $\Gamma \geq C(\Lambda)$. Therefore, there is a $y \in [\omega_\varepsilon^+(.,0) \leq 0]$ such that $\delta_\varepsilon^+(x,0) = d(x,y) > 0$. This yields

$$\begin{aligned} \frac{\Gamma}{2}\varepsilon &\leq \omega_\varepsilon^+(x,0) - \omega_\varepsilon^+(y,0) = \omega(x,0) - \omega(y,0) \\ &= \delta_0(x) - \delta_0(y) \leq C(\Lambda)|x-y| \leq C(\Lambda)d(x,y) = C(\Lambda)\delta_\varepsilon^+(x,0). \end{aligned}$$

We conclude

$$\lambda_\varepsilon^+(x,0) = \frac{\delta_\varepsilon^+(x,0)}{\varepsilon} - \pi - \alpha \geq c_0(\Lambda)\Gamma - \pi - \alpha.$$

Since $\alpha \leq C(\Lambda)$, we can choose $\Gamma \geq C(\Lambda)$ to get

$$\lambda_\varepsilon^+(x,0) \geq \frac{\pi}{2};$$

hence

$$\psi_\varepsilon^+(x,0) = 1 \geq \varphi_\varepsilon(x,0),$$

establishing (89) and therefore (88).

We take (x_0,t_0) with $\omega(x_0,t_0) < 0$. There is $\tau > 0$ and a neighborhood $U(x_0,t_0)$ such that

$$\omega \leq -\tau \quad \text{on } U(x_0,t_0);$$

hence

$$\omega_\varepsilon^+ \leq -\frac{\tau}{2} \quad \text{on } U(x_0,t_0) \quad \text{for } 0 < \varepsilon < \varepsilon_0(\Lambda,\tau).$$

On $U(x_0,t_0)$, we have

$$\delta_\varepsilon^+ = 0;$$

hence

$$\lambda_\varepsilon^+ \leq -\pi$$

and, finally,

$$\varphi_\varepsilon \leq \psi_\varepsilon^+ = -1 \quad \text{on } U(x_0,t_0). \qquad \square$$

Remark 4.6. When $H^n([\omega = 0]) = 0$, then the limit of φ_ε is uniquely determined in $L^1(\mathbb{R}^n \times [0,T[)$ by Theorem 4.5.

When fattening occurs—that is, $H^n([\omega = 0]) > 0$—there remains an ambiguity.

Acknowledgments. We thank Maurizio Paolini and Mete Soner for several discussions related to the topic of this article.

REFERENCES

[1] S. Allen and J. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica, 27 (1979), pp. 1084–1095.

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

[3] G. Belletini and M. Paolini, *Quasi-optimal error estimates for the mean-curvature flow with forcing term*, Differential Integral Equations, 8 (1995), pp. 735–752.

[4] G. Belletini and M. Paolini, *Anisotropic motion by mean curvature in the context of Finsler geometry*, Quaderno 49, Universitá degli Studi di Milano, Milan, Italy, 1994; Hokkaido Math. J., to appear.

[5] J. F. Blowey and C. M. Elliott, *Curvature dependent phase boundary motion and parabolic obstacle problems*, in Proc. IMA Workshop on Degenerate Diffusion, Vol. 47, W. Ni, L. Peletier, and J. L. Vasquez, eds., Springer-Verlag, Berlin, 1993, pp. 19–60.

[6] X. Chen, *Generation and propagation of interface in reaction-diffusion equations*, J. Differential Equations, 96 (1992), pp. 116–141.

[7] Y. 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

[8] X. Chen and C. M. Elliott, *Asymptotics for a parabolic double obstacle problem*, Proc. Roy. Soc. London Ser. A, 444 (1994), pp. 429–445.

[9] M. G. Crandall, H. Ishi, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[10] P. de Mottoni and M. Schatzmann, *Evolution géométrique d'interfaces*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 453–458.

[11] C. M. Elliott and R. Schätzle, *The limit of the anisotropic double-obstacle Allen–Cahn equation*, Proc. Roy. Soc. Edinburgh Sect. A, Vol. 126, to appear.

[12] C. M. Elliott, M. Paolini, and R. Schätzle, *The limit of the anisotropic double-obstacle Allen–Cahn equation with kinetic term*, Math. Models Methods Appl. Sci., to appear.

[13] L .C. Evans, H. M. Soner, and P. E. Souganidis, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., XLV (1992), pp. 1097–1123.

[14] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs 23, AMS, Providence, RI, 1968.

[15] G. B. McFadden, A. A. Wheeler, R. J. Braun, S. R. Coriell, and R. F. Sekerka, *Phase-field models for anisotropic interfaces*, Phys. Rev. E, 48 (1993), pp. 2016–2024.

[16] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[17] R. H. Nochetto, M. Paolini, and C. Verdi, *Sharp error analysis for curvature dependent evolving fronts*, Math. Models Methods Appl. Sci., 3 (1993), pp. 711–723.

[18] R. H. Nochetto and C. Verdi, *Convergence of double obstacle problems to the geometric motions of fronts*, Quaderno 39, Universitá degli Studi di Milano, Milan, Italy, 1994.

[19] H. M. Soner, *Motion of a set by the curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.

[20] J. E. Taylor and J. W. Cahn, *Linking anisotropic sharp and diffuse surface motion laws via gradient flows*, J. Statist. Phys., 77 (1993), pp. 183–197.

[21] A. A. Wheeler and G. B. McFadden, *A ξ-vector formulation of anisotropic phase-field models: 3-D asymptotics*, Mathematics Research Report AM-94-05, University of Bristol, Bristol, UK, 1994.

SIAM J. Math. Anal.
Vol. 28, No. 2, pp. 304–321, March 1997

004

THE ASYMPTOTIC BEHAVIOR OF THE HYPERBOLIC CONSERVATION LAWS WITH RELAXATION ON THE QUARTER-PLANE[*]

SHINYA NISHIBATA[†] AND SHIH-HSIEN YU[‡]

Abstract. The hyperbolic conservation laws with relaxation appear in many physical models such as those for gas dynamics with thermo-nonequilibrium, elasticity with memory, flood flow with friction, and traffic flow. The main concern of this article is the long-time behavior of the interaction between the relaxations and the boundary conditions. In this article, we investigate this problem for a simple model of a 2×2 system. It is proven that the solution of the system asymptotically converges to a traveling wave moving away from the boundary under suitable conditions on the boundary.

Key words. equilibrium state, subcharacteristic condition, boundary condition

AMS subject classification. 35

PII. S0036141095276506

1. Introduction.

The phenomena of relaxation are present in the kinetic theory of gas, e.g., the Broadwell model [1], [2], [8], [13], elasticity with memory, gas flow with thermo-nonequilibrium, water waves, etc. Liu [9] gave a 2×2 strictly hyperbolic system as a model equation for the relaxation phenomenon, where he studied the asymptotic behavior of the solution for both rarefaction waves and traveling waves. In the same paper, the validity of the Chapman–Enskog expansion was also investigated. In most of the physical situations where relaxations occur, it is inevitable to take boundary effects into account. In this paper, we would like to extend the results of Liu [9] to cases where the boundary effects are taken into consideration.

In the case of the initial value problem, the asymptotic behavior of Liu's 2×2 system is governed by an equilibrium equation. However, for the initial-boundary value problem in the first quadrant, we must consider the effect of the boundary condition. The number of boundary conditions required for the equilibrium equation is one or zero depending on whether the equilibrium characteristic speed is positive or not. For the 2×2 system, the number of the boundary conditions required is the same as the number of characteristics out of the boundary, which could be zero, one, or two depending on the directions of the characteristics. Therefore, for the 2×2 system and for the equilibrium equation, they may not require the same number of boundary conditions. In the case where the numbers are different, there might be some boundary layer appearing; cf. [8] and [15]. In [8], there are various interesting behaviors due to the boundary conditions. In this paper, we consider the case when both the 2×2 system and the equilibrium equation require one boundary condition.

Previously, Chen [3] studied the nonlinear diffusion wave for Liu's model. Nishibata [12] followed Liu's model equation to study the stability of the stationary solution for the initial-boundary problems, which is closely related to the work in this paper.

[*] Received by the editors September 25, 1995; accepted for publication (in revised form) November 7, 1995. This research was performed while the authors were in the Ph.D. program at Stanford University, Stanford, CA.

http://www.siam.org/journals/sima/28-2/27650.html

[†] Department of Mathematics, Stanford University, Stanford, CA 94305-2125. Current address: Fukuoka Institute of Technology, Fukuoka 811-02, Japan (shinya@fit.ac.jp).

[‡] Department of Mathematics, Stanford University, Stanford, CA 94305-2125. Current address: Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90025-1555 (shyu@math.ucla.edu).

Consider the following system of two quasi-linear linear hyperbolic equations as a model of relaxation [9]:

$$\left\{\begin{array}{l} u_t + f(u,v)_x = 0 \quad \text{for } x, t > 0, \\ v_t + g(u,v)_x = h(u,v). \end{array}\right. \tag{1.1}$$

The first equation represents a conservation law for and the second equation represents a rate equation. Suppose that this system is a strictly hyperbolic; $\lambda_1(u,v) < \lambda_2(u,v)$, where λ_1 and λ_2 are the eigenvalues of the Jacobian matrix

$$\begin{pmatrix} f_u(u,v) & f_v(u,v) \\ g_u(u,v) & g_v(u,v) \end{pmatrix}.$$

In order to make the conservation law and the rate equation strongly coupled, we assume that

$$f_v(u,v) \neq 0 \quad \text{for all } (u,v) \text{ under consideration.}$$

The term $h(u,v)$ acts as the a source (or a sink) when v is less (or greater) than the equilibrium state $v_*(u)$. $h(u,v)$ often assumes the form

$$h(u,v) = \frac{v_*(u) - v}{\tilde{\tau}(u)}$$

for some positive function $\tilde{\tau}(u)$, the relaxation time. We make the general assumption

$$\frac{\partial h(u,v)}{\partial v} < 0, \quad h(u, v_*(u)) = 0,$$

for all (u,v) under consideration. We impose boundary values on u:

$$u(0,t) = u_-, \qquad u(\infty,t) = u_+, \tag{1.2}$$

where u_- and v_- are constants. Since the solution at $x = \infty$ is permanent, it is an equilibrium state:

$$v(\infty,t) = v_*(u_-). \tag{1.3}$$

When the solution is close to the equilibrium state, we often ignore the rate equation and replace the conservation law with the equilibrium equation:

$$\frac{\partial u}{\partial t} + \frac{\partial f_*(u)}{\partial x} = 0, \quad f_*(u) =: f(u, v_*(u)). \tag{1.4}$$

When $(u_-, v_*(u_-)$ and $(u_+, v_*(u_+))$ are connected by a shock wave for the system (1.1), we have the Rankine–Hugoniot relation with the shock speed σ:

$$\sigma = \frac{f_*(u_-) - f_*(u_+)}{u_- - u_+}.$$

For the stability of (1.1), we require the subcharacterisitc condition $\lambda_1 < \sigma < \lambda_2$. Here we are interested in the case $\sigma > 0$. In order to make the equilibrium equation (1.4) and system (1.1) require the same number of boundary conditions on the boundary $x = 0$, we assume the following:

$$\lambda_1(u,v) < 0 < \sigma < \lambda_2(u,v) \quad \text{for all } (u,v) \text{ under consideration.} \tag{1.5}$$

Furthermore, for simplicity, we assume that

$$f_*^{''}(u) > 0. \tag{1.6}$$

We define $\hat{u}$ as the number uniquely determined to satisfy

$$f_*'(\hat{u}) = \sigma.$$

We denote by $(\phi(x-\sigma t), \psi(x-\sigma t))$ a traveling-wave solution of the Cauchy problem of system (1.1) that connects the end states $(u_-, v_*(u_-))$ and $(u_+, v_*(u_+))$. When $|u_- - u_+|$ tends to 0, the curve $L = \{(\phi(x), \psi(x)) : x \in \mathbf{R}\}$ degenerates to the point $(\hat{u}, v_*(\hat{u}))$ by (1.6), i.e., when $|u_- - u_+|$ is sufficiently small,

$$\|\phi(x) - \hat{u}\|_\infty + \|\psi(x) - v_*(\hat{u})\|_\infty \le K_0 |u_- - u_+|. \tag{1.7}$$

where K_0 is an constant depending on $\hat{u}$ only (see [9]).

Henceforth, we normalize the traveling wave as follows:

$$\phi(0) = \hat{u}.$$

Remark. In the remainder of this paper, all of the constants depend only on $\nabla^i f(\hat{u}, \hat{v})$, $\nabla^i g(\hat{u}, \hat{v})$, and $\nabla^i h(\hat{u}, \hat{v})$ $(i = 0, 1, 2)$, where $\hat{v} = v_*(\hat{u})$ unless otherwise mentioned.

The initial conditions are chosen so that $(u(x,0), v(x,0)) = (u_0(x), v_0(x))$ with $u_0,\ v_0 \in C^3[0,\infty)$ and that the following is satisfied:

$$\begin{aligned} &\sum_{i=0}^{3} |\nabla^i (u_0(x) - \phi(x - x_0))|^2 + \sum_{i=0}^{3} |\nabla^i (v_0(x) - \psi(x - x_0))|^2 \\ &\le \begin{cases} \delta_1 |\phi(x - x_0) - u_+| & \text{if } x > x_0, \\ \delta_1 |\phi(x - x_0) - u_-| & \text{if } 0 < x < x_0, \end{cases} \end{aligned} \tag{1.8}$$

where δ_1 and x_0 are certain constants to be determined later.

Recall that we set the boundary conditions in (1.2). In addition, suppose that λ_1 and λ_2 satisfy the following condition:

$$\begin{gathered} -\lambda_2(u,v) < \lambda_1(u,v) < 0 < -\frac{(\lambda_1 + \lambda_2)}{\lambda_1 \lambda_2(u,v)} < \sigma < \lambda_2(u,v) \\ \text{for all } (u,v) \text{ under consideration} \end{gathered} \tag{1.9}$$

This additional assumption will later be necessary to estimate the rate of convergence of $v(0,t) - v_*(u_-)$ on the boundary.

We introduce the perturbations $\bar{u}$ and $\bar{v}$:

$$\begin{cases} u(x,t) = \phi(x - \sigma t - x_0) + \bar{u}(x,t), \\ v(x,t) = \phi(x - \sigma t - x_0) + \bar{v}(x,t). \end{cases} \tag{1.10}$$

Note that in the rest of this section, (ϕ, ψ) will stand for $(\phi(x-\sigma t-x_0), \psi(x-\sigma t-x_0))$ unless otherwise mentioned.

From (1.1), we have the system of differential equations for the perturbation $(\bar{u}, \bar{v})$:

$$\begin{cases} \bar{u}_t + (\nabla f \cdot (\bar{u}, \bar{v}))_x = -M(f)_x, \\ \bar{v}_t + (\nabla g \cdot (\bar{u}, \bar{v}))_x = \nabla h \cdot (\bar{u}, \bar{v}) - M(g)_x + M(h), \end{cases} \tag{1.11}$$

where $M(f) =: f(\phi+\bar{u},\psi+\bar{v}) - f - \nabla f\cdot(\bar{u},\bar{v})$ for any function f, where f, ∇f, ∇g, and ∇h are evaluated at $(\phi(x-\sigma t-x_0),\psi(x-\sigma t-x_0))$.

Remark. The function $M(f)$ is a function of higher than second order in $(\bar{u},\bar{v})$.

THEOREM 1.1 (local existence theorem). *For system* (1.1)*, there is a* $\delta_3 > 0$ *such that for any* $\sum_{i=0}^{1}\left(\|\partial_x^i(u(\cdot,0)-\hat{u})\|_\infty + \|\partial_x^i(v(\cdot,0)-v_*(\hat{u}))\|_\infty\right) \le \delta_3/2$*, there is constant* $\tau(\delta_3) > 0$ *such that the solution* (u,v) *exists in* $[0,\infty)\times[0,\tau(\delta_3))$ *and satisfies*

$$\sup_{0<t<\tau(\delta_3)}\left(\sum_{i=0}^{1}\|\partial_x^i(u(\cdot,t)-\hat{u})\|_\infty + \|\partial_x^i(v(\cdot,t)-v_*(\hat{u}))\|_\infty\right) \le \delta_3.$$

The proof is given by the standard iteration method (see [7]).

We denote

$$N(t) =: \|\phi'\|_\infty + \|\psi'\|_\infty + \mathrm{sup}_{0<\tau<t}\left(\sum_{i=0}^{1}\|\partial_x^i\bar{u}(\cdot,\tau)\|_\infty + \sum_{i=0}^{1}\|\partial_x\bar{v}(\cdot,\tau)\|_\infty\right),$$
$$t_n =: n\tau(\delta_2), \quad n\in\mathbf{N}.$$

LEMMA 1.2 (main lemma). *For given small* δ_0 *and* δ_2 *satisfying* $|u_- - u_+| < \delta_0 < \delta_2 \ll \delta_3$*, we choose* δ_1 *and* x_0 *so that* δ_1*,* $e^{-|u_- - u_+|x_0}$*, and* $\delta_1 x_0$ *are sufficiently small. Suppose that* $N(t_{n-1}) < \delta_2/2$ *holds. Then we have* $N(t_n) < \delta_2/2$*.*

Proof. See section 3.

THEOREM 1.3 (global existence theorem). *For given small* δ_0 *and* δ_2 *satisfying* $|u_- - u_+| < \delta_0 < \delta_2 \ll \delta_3$*, we choose* δ_1 *and* x_0 *to satisfy the requirement that* δ_1*,* $e^{-|u_- - u_+|x_0}$*, and* $\delta_1 x_0$ *are sufficiently small. Then the solution* $(u(x,t),v(x,t))$ *exists globally and satisfies*

$$\sup_{t>0}\Bigg(\|\phi'\|_\infty + \|\psi'\|_\infty + |u_- - u_+| + \sum_{i=0}^{1}\|\partial_x^i(u(\cdot,t)-\phi(\cdot-\sigma t-x_0))\|_\infty$$
$$+\sum_{i=0}^{1}\|\partial_x^i(v(\cdot,t)-\psi(\cdot-\sigma t-x_0))\|_\infty\Bigg) < \frac{\delta_2}{2}.$$

Proof. From Theorem 1.1 and Lemma 1.2, Theorem 1.3 follows.

THEOREM 1.4 (the convergence theorem). *Given that* $|u_- - u_+| < \delta_0 < \delta_2 \ll \delta_3$ *are small, suppose that* δ_1*,* $e^{-|u_- - u_+|x_0}$*, and* $\delta_1 x_0$ *are sufficiently small. Then there is a unique* $s_\infty < \infty$ *such that* $(u(x,t),v(x,t))$ *converges to the traveling wave* $(\phi(x-\sigma t-s_\infty),\psi(x-\sigma t-s_\infty))$ *when* t *tends to* ∞*.*

Proof. See section 3.

The difficulty in the present problem is due to the fact that the wave is moving away from the boundary, so we cannot determine the location error. In section 2, this difficulty is resolved by introducing the Riemann invariants and a suitable weight function to obtain an a priori estimate for the rate of $v(0,t) - v_*(u_-)$ converging to 0.

The same difficulty in determining the location error also arises in the initial-boundary problem for viscous conservation laws and the Cauchy problem for undercompressible flow. For the initial-boundary problems for the Burgers equation, Yu [14] and Liu and Yu [10] used the boundary gradient estimate to obtain the location error and get the asymptotic stability. In case of undercompressible flow, Zumbrum and Liu [11] used the pointwise estimate to overcome the difficulty of determining the location error.

In section 3, we introduce a sequence of the location error associated with a time sequence tending to a infinity. By using the sequence of the location error and the decreasing rate of $|v(0,t) - v_*(u_-)|$, the standard energy method can be applied to show the global existence. We also show that the solution asymptotically converges to the traveling wave.

2. Initial-boundary problem and boundary-value estimate.

2.1. Convergence of boundary value. We introduce new variables r_1 and r_2 to diagonalize the linear parts of (1.11).

Notation.

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} := L(x,t) \cdot \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} \tag{2.1}$$

with $L(x,t) := R(x,t)^{-1}$, where

$$R(x,t) := \begin{pmatrix} -f_v & -f_v \\ f_u - \lambda_1 & f_u - \lambda_2 \end{pmatrix}.$$

Substitute the variables r_1 and r_2 into (1.1); it then follows that

$$\begin{aligned} &\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}_t + \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}_x - L(x,t) \begin{pmatrix} 0 \\ \nabla h(\bar{u},\bar{v}) \end{pmatrix} \\ &\quad = L(x,t)_t R(x,t) \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} - \begin{pmatrix} \lambda_1 & 0 \\ 0, & \lambda_2 \end{pmatrix} L(x,t) R(x,t)_x \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \\ &\qquad + L(x,t) \begin{pmatrix} 0 \\ M(h) \end{pmatrix} + L(x,t) \begin{pmatrix} -M(f)_x \\ -M(g)_x \end{pmatrix}. \end{aligned} \tag{2.2}$$

Next, we set

$$\begin{aligned} lW(x,t) &:= e^{\frac{2H\cdot(x-\sigma t)}{c_2 - c_1}}, \\ \mu(u,v) &:= f_u(u,v) - (f_v h_u h_v^{-1})(u,v), \\ H &:= h_v(\hat{u},\hat{v}) < 0, \\ c_i &=: \lambda_i(\hat{u},\hat{v}) \quad \text{with } i = 1,2, \\ b &= \frac{(c_2 - \sigma)}{(\sigma - c_1)}, \end{aligned}$$

where the variable $\mu(u,v)$ is the dynamic characteristic speed defined in [9]. From the definition of W, it follows that

$$\frac{W_x}{W} = \frac{2H}{(c_2 - c_1)}. \tag{2.3}$$

We multiply (2.2) by $W(x,t)(r_1, b\ r_2)$ and integrate it over $0 < x < \infty$. We substitute (1.7) into this integration. We then substitute (2.3) and Schwartz's inequality into this integration. Then there is a positive constant K_1 such that

(2.4)

$$\frac{d}{dt}\int_0^\infty W \cdot (r_1^2 + b\ r_2^2)(x,t)dx + ((-\lambda_1 r_1^2 - b\ \lambda_2 r_2^2) \cdot W)(0,t)$$

$$\leq \frac{-kH}{c_2 - c_1} \int_0^\infty W \cdot (r_1^2 + b\ r_2^2) dx$$
$$+ K_1 \left(N(t)(r_1^2(0,t) + b\ r_2^2(0,t)) + |u_- - u_+| \int_0^\infty W \cdot (r_1^2 + b\ r_2^2)(x,t) dx \right),$$

where

$$k = -2(c_2 - \sigma) + \sqrt{4(c_2 - \sigma)^2 + \frac{((c_2 - \sigma) + (\sigma - c_1) b)^2}{b}}.$$

Next, we compute the terms of (2.4) evaluated at $x = 0$. Substituting $x = 0$ into (1.10), it follows that

$$\bar{u}(0,t) = u_- - \phi(-x_0 - \sigma t). \tag{2.5}$$

From (2.5) and (2.1), we have that

$$-f_v\ (r_1 + r_2)(0,t) = u_- - \phi(-\sigma t - x_0). \tag{2.6}$$

From (2.6), we have that

$$r_2^2(0,t) = \left(\frac{-(u_- - \phi(-x_0 - \sigma t))}{f_v} - r_1(0,t) \right)^2. \tag{2.7}$$

Substituting (2.7) in the boundary term on the left-hand side (LHS) of (2.4), we obtain

(2.8)
$$-(\lambda_1 r_1^2 + \lambda_2 b\ r_2^2)(0,t) = -\left(\lambda_1 r_1^2 + b\ \lambda_2 \left(r_1 + \frac{(u_- - \phi(-\sigma t - x_0))}{f_v} \right)^2 \right)$$
$$> -\frac{(\lambda_1 + b\ \lambda_2) r_1^2(0,t)}{2} - K_0 \lambda_2 \left(\frac{(u_- - \phi(-x_0 - \sigma t)}{f_v} \right)^2,$$

where K_0 is a positive constant depending only on $\hat{u}$. From assumption (1.9), $-(\lambda_1 + b\lambda_2) > 0$. Hence if $N(t)K_1 \ll -(\lambda_1 + b\ \lambda_2)/8$, then the boundary term on the right-hand side (RHS) of (2.4) can be absorbed into the positive boundary term on the LHS of (2.4). Therefore, there are constants K_2 and K_1 such that

(2.9)
$$\frac{d}{dt} \int_0^\infty W(r_1^2 + b\ r_2^2) dx \leq \frac{-H(k + K_1(N(t) + |u_- - u_+|))}{c_2 - c_1} \int_0^\infty W(r_1^2 + b\ r_2^2) dx$$
$$+ K_2(\phi(x_0 - \sigma t) - u_-)^2 W(0,t).$$

Lemma 2.1.

$$d := -2c_2 + \sqrt{4(c_2 - \sigma)^2 + \frac{((c_2 - \sigma) + (\sigma - c_1) b)^2}{b}} < 0.$$

Proof. Due to an algebraic manipulation of (1.9), Lemma 2.1 follows.

LEMMA 2.2. *If $K_1(|u_- - u_+| + N(t)) < -d/4$, then for $t' \in (0,t)$,*

$$\int_0^1 \left(r_1^2(x,t') + b\ r_2^2(x,t')\right) dx \le e^{\frac{-dHt'}{(c_2-c_1)}} \int_0^\infty (r_1^2(x,0) + b\ r_2^2(x,0))W(x,0)dx$$
$$+ K_2 e^{\frac{-2H}{(c_2-c_1)}} \int_0^{t'} e^{\frac{-dH(t'-s)}{2(c_2-c_1)}} (\phi(-x_0 - \sigma s) - u_-)^2 ds. \tag{2.10}$$

Proof. From (2.9), we have that for $t' \in (0,t)$,

$$\int_0^1 (r_1^2(x,t') + b\ r_2^2(x,t'))W(1,t')dx < \int_0^\infty \left((r_1^2 + b\ r_2^2)W\right)(x,t')dx$$
$$\le e^{\frac{-H(k+K_1(N(t')+|u_- - u_+|))t'}{c_2-c_1}} \int_0^\infty \left((r_1^2 + b\ r_2^2)W)\right)(x,0)dx \tag{2.11}$$
$$+ \int_0^{t'} K_2(\phi(-\sigma s - x_0) - u_-)^2 W(0,s) e^{\frac{-H(k+K_1(N(t')+|u_- - u_+|))(t'-s)}{c_2-c_1}} ds.$$

Dividing both sides of (2.11) by $W(1,t')$ and substituting the condition $K_1(|u_- - u_+| + N(t)) < -d/4$ into (2.11), Lemma 2.2 follows.

THEOREM 2.3. *Suppose for a given $\delta_2\,(< -d/4K_1)$ that the following holds:*

$$\text{if } |u_- - u_+| + N(t) \le \delta_2, \quad |u_- - u_+| \ll 1, \quad \text{and} \quad x_0 \gg 1, \tag{2.12}$$
then there is a constant K_3 such that for $t' \in (0,t)$,

$$|v(0,t') - v_*(u_-)| \tag{2.13}$$
$$\le K_3 \left\{\delta_2^{1/2}\delta_1^{1/4} + \delta_1^{1/2}\right\} \left[(\phi(-x_0 - \sigma t') - u_-)^2 + |\phi(-x_0/2) - u_-| e^{dt'/2}\right]^{1/4}_.$$

Proof. By the condition on the initial value ((1.8)) and Lemma 2.2, we have

$$\int_0^\infty W(x,0)(\bar{u}^2(x,0) + \bar{v}^2(x,0))dx$$
$$= O(1)\delta_1 \left(\int_{[0,x_0/2]} + \int_{[x_0/2,x_0]}\right) W(x,0)|\phi(x - x_0) - u_-|^2 dx$$
$$+ O(1)\delta_1 \int_{x_0}^\infty W(x,0)|\phi(x - x_0) - u_+|^2 dx \tag{2.14}$$
$$= O(1)\delta_1 \left\{|\phi(-x_0/2) - u_-| + e^{\frac{Hx_0}{c_2-c_1}}\right\}.$$

Hence from (2.14) and Lemma 2.2, we have

$$\int_0^1 \bar{u}^2(x,t') + \bar{v}^2(x,t')dx \tag{2.15}$$
$$= O(1)\delta_1 \left\{|\phi\left(\frac{-x_0}{2}\right) - u_-| e^{dt'/2} + \left(\phi\left(\frac{-x_0 - \sigma t'}{2}\right) - u_-\right)^2\right\}.$$

From $N(t) + |u_- - u_+| < \delta_2$, we have that for $t' \in (0,t)$,

$$\int_0^1 |\bar{u}_x(x,t')|^2 + |\bar{v}_x(x,t')|^2 dx \le N(t)^2 \le \delta_2^2. \tag{2.16}$$

From (2.16) and (2.15),

$$\text{(2.17)}\quad \sup_{0\le x,x'\le 1} |v(x,t') - v(x',t')| \le K_3'\delta_2^{1/2}\delta_1^{1/4}\left[(\phi(-x_0-\sigma t') - u_-)^2 + |\phi(-x_0/2) - u_-|e^{dt'/2}\right]^{1/4},$$

and from (2.15), there is an $\bar{x} \in [0,1]$ satisfying

$$\text{(2.18)}\quad |v(\bar{x},t')| \le \left\{K_3'\delta_1\left(\left|\phi\left(-\frac{x_0}{2}\right) - u_-\right| e^{dt'/2} + \left(\phi\left(-\frac{x_0-\sigma t'}{2}\right) - u_-\right)^2\right)\right\}^{1/2},$$

where K_3' is a constant depending only on $\hat{u}$. From (2.17) and (2.18), Theorem 2.3 follows. Here note that by standard ordinary differential equation theory, the first term of the RHS of (2.13) converges to zero with an exponential rate. Hence $v(0,t)$ converges to $v_*(\hat{u})$ with an exponential rate.

2.2. The sequence of the location error. From the a priori boundary estimate Theorem 2.3, we obtain that $v(0,t)$ converges to $v_*(\hat{u})$ with an exponential rate. Once we get the rate of convergence, we can handle the boundary terms that appear in the usual energy estimate. However, we still have to deal with the asymptotic state caused by the location error, which is thus far unexamined. In order to resolve this difficulty, we introduce a sequence of the location error.

DEFINITION 2.4. *The sequence of the location error* $\{\sigma t + s_n\}_{n\in\mathbf{N}}$ *is associated with the time sequence* $\{t_n | t_n = n\tau(\delta_2),\ n \in \mathbf{N}\}$ (*cf. the comment below Theorem* 1.1). *We define* $\{s_n\}_{n\in\mathbf{N}}$ *by the following implicit equation:*

$$\text{(2.19)}\qquad \int_0^\infty u(x,t_n) - \phi(x-\sigma t_n - s_n)dx = 0.$$

Here s_n *is uniquely determined because* ϕ *is monotonic* (*see Liu* [9]).

From (2.19), we have the following identity due to the conservation laws:

$$\text{(2.20)}$$
$$\begin{aligned}
0 &= \int_0^\infty u(x,t_n) - \phi(x-s_n-\sigma t_n)dx \\
&= \int_0^\infty u(x,t_n) - \phi(x-x_0-\sigma t_n)dx + \int_0^\infty \phi(x-x_0-\sigma t_n) - \phi(x-s_n-\sigma t_n)dx \\
&= \int_0^\infty u(x,0) - \phi(x-x_0)dx + \int_0^{t_n} \frac{d}{dt}\left\{\int_0^\infty u(x,t) - \phi(x-x_0-\sigma t)dx\right\}dt \\
&\qquad + \int_0^\infty \phi(x-x_0-\sigma t_n) - \phi(x-s_n-\sigma t_n)dx \\
&= \int_0^{t_n} f(u_-, v(0,t)) - f(\phi(-x_0-\sigma t), \psi(-x_0-\sigma t))dt + \int_0^\infty u(x,0) - \phi(x-x_0)dx \\
&\qquad + \int_{x_0}^{s_n} (\phi(-\sigma t_n - s) - u_+)ds \\
&= O(1)\int_0^{t_n} |\psi(-x_0-\sigma t) - v_*(u_-)| + |v(0,t) - v_*(u_-)|dt
\end{aligned}$$

$$
\begin{aligned}
&+O(1)\left\{\int_0^{t_n}|\phi(-x_0-\sigma t)-v_*(u_-)|dt+|u_--u_+||x_0-s_n|\right\}\\
&+\int_0^\infty u(x,0)-\phi(x-x_0)dx.
\end{aligned}
$$

Hence we have that when $|u_--u_+|\ll 1$, it follows from Theorem 2.3 that

$$
|s_n-x_0|<\frac{O(1)\left(\delta_1+\delta_1^{1/4}\delta_2^{1/2}\left|\phi\left(-\frac{x_0}{4}\right)-u_-\right|\right)}{|u_--u_+|^3}. \tag{2.21}
$$

3. Proof of Lemma 1.2. We will split this section into five subsections in order to prove the main lemma. Then in the last subsection, we will show that the wave (u,v) converges to the asymptotic wave $(\phi(x-\sigma t-s_\infty),\psi(x-\sigma t-s_\infty))$, where the asymptotic location error $\sigma t+s_\infty$ is defined formally as

$$
s_\infty=\lim_{n\to\infty}s_n. \tag{3.1}
$$

The existence of the above limit will be shown at the end of this paper. In the next four subsections, we will show that

$$
N(t_n)+|u_--u_+|\le\frac{\delta_2}{2}\quad\text{for } n\in\mathbf{N} \tag{3.2}
$$

by induction.

3.1. Preliminaries for the global existence theorem (Theorem 1.3). Because of the assumption of Lemma 1.2, $N(t_0)+|u_--u_+|\le\delta_2/2$ holds for $t_0=0$. Hence we assume that the solution (u,v) exists in $[0,t_{n-1}]$ and that (3.2) holds for $n-1$. Then by the local existence theorem (Theorem 1.1), the existence of the solution (u,v) can be extended to $[0,\infty)\times[0,t_n]$, and it satisfies

$$
N(t_n)+|u_--u_+|<\delta_2.
$$

We define the function $z(x,t)$:

$$
z(x,t)=-\int_x^\infty u(y,t)-\phi(y-\sigma t-s_n)dy. \tag{3.3}
$$

Then from Theorem 2.3, we have the following for $0<t\le t_n$:

$$
\begin{aligned}
&|z(0,t)|= \\
&\frac{O(1)\left\{\delta_2^{1/2}\delta_1^{1/4}\left((\phi(-x_0-\sigma t)-u_-)^2+\left|\phi\left(-\frac{x_0}{2}\right)-u_-\right|e^{dt/2}\right)^{1/4}+|\phi(x_0-\sigma t)-u_-|\right\}}{|u_--u_+|},
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
&|z_t(0,t)|= \\
&O(1)\left\{\delta_2^{1/2}\delta_1^{1/4}\left((\phi(-x_0-\sigma t)-u_-)^2+\left|\phi\left(-\frac{x_0}{2}\right)-u_-\right|e^{dt/2}\right)^{1/4}+|\phi(x_0-\sigma t)-u_-|\right\}.
\end{aligned} \tag{3.5}
$$

Define $\tilde u$ and $\tilde v$ as follows:

$$
u(x,t)=\phi(x-\sigma t-s_n)+\tilde u(x,t),\qquad v(x,t)=\psi(x-\sigma t-s_n)+\tilde v(x,t).
$$

$$z_x(0,t) = u_- - \phi(-\sigma t - s_n).$$

Substitute the function $z(x,t)$ in (1.1). From the conservation law in (1.1), we have that

$$\begin{aligned} &z_t + f(\phi(x-\sigma t - s_n) + z_x, \psi(x - \sigma t - s_n) + \tilde{v})) \\ &\quad - f(\phi(x-\sigma t - s_n), \psi(x-\sigma t - s_n)) = 0. \end{aligned} \tag{3.6}$$

In the next three subsections, the functions f, g, ∇f, ∇g, λ_1, λ_2, and h are evaluated at $(\phi(x-\sigma t - s_n), \psi(x - \sigma t - s_n))$ unless otherwise specified. Therefore,

$$\tilde{v}(x,t) = -f_v^{-1}(z_t + f_u z_x + Q_0(f)), \tag{3.7}$$

where $Q_0(f) = f(z_x + \phi, \tilde{v} + \psi) - f(\phi,\psi) - \nabla f(z_x, \tilde{v})$. Hence $Q_0(f) = O(1)(z_x^2 + \tilde{v}^2)$. Substitute (3.7) into the rate equation in (1.1); then it follows that

$$\begin{aligned} &z_{tt} + (\lambda_1 + \lambda_2) z_{xt} + \lambda_1 \lambda_2 z_{xx} - h_v(z_t + \mu z_x) \qquad (3.8) \\ &\quad = f_v^{-1}\left(f_{vt} + g_v f_{vx} - g_{vx}\right)(z_t + f_u z_x) + Q_0(f)(f_v^{-1} g_v f_{vx} - g_{vx} + h_v) \\ &\qquad - f_v Q_0(h) - Q_0(f)_t - g_v Q_0(f)_x + f_v Q_0(g)_x \qquad (3.9) \\ &\quad := (\text{RHS})_{3.8} \\ &\quad =: f_v^{-1}\left(f_{vt} + g_v f_{vx} - g_{vx}\right)(z_t + f_u z_x) + (\text{RHS})'_{(3.8)}. \end{aligned}$$

Without loss of generality, we may assume that

$$h_v(\hat{u}, v_*(\hat{u})) = -2 \tag{3.10}$$

by rescaling the coordinates (x,t). This condition is given simply for convenience of calculation.

3.2. Step I: The telegraph equation. Before we carry out the calculation for the nonlinear equation (1.1), we introduce the linear telegraph equation in order to arrange the nonlinear terms in (1.1). By assuming for the moment that in equation (3.8), λ_1, λ_2, and μ are all constant and the RHS is identically zero, we have

$$\begin{cases} \xi_{tt} + (c_1 + c_2)\xi_{xt} + c_1 c_2 \xi_{xx} + 2(\xi_t + \sigma \xi_x) = 0 \quad \text{for } x, t > 0, \\ c_1 < 0 < \sigma < c_2, \end{cases} \tag{3.11}$$

Integrate $(3.11)\xi + (3.11)\xi_t + (3.11)((c_1 + c_2)/2)\xi_x$ over $0 < x < \infty$, $0 < \tau < t$. We obtain that

(3.12)

$$\begin{aligned} &\int_0^\infty \left(-\frac{c_1 c_2}{2}\xi_x^2 + \frac{(\xi_t + (c_1 + c_2)\xi_x)^2}{4} + \frac{(\xi_t + 2\xi)^2}{4}\right)(x,t)dx - \frac{(c_1 + c_2)}{2}\xi^2(0,t) \\ &\qquad + \int_0^t \left(-\sigma \xi^2 - c_1 c_2 \xi_x \xi - \frac{(c_1 + c_2)}{4} c_1 c_2 \xi_x \xi_t - \frac{(c_1 + c_2)}{2}\xi_t^2\right)(0,t)d\tau \\ &\qquad + \int_0^\tau \int_0^\infty \xi_t^2 + 2\sigma \xi_x \xi_t + (\sigma(c_1 + c_2) - c_1 c_2)\xi_x^2 dx d\tau \\ &= \int_0^\infty \left(-\frac{c_1 c_2}{2}\xi_x^2 + \frac{(\xi_t + (c_1 + c_2)\xi_x)^2}{4} + \frac{(\xi_t + 2\xi)^2}{4}\right)(x,0)dx - \frac{c_1 + c_2}{2}\xi^2(0,0). \end{aligned}$$

The integrand in the double integral in (3.12) is positive due to the subcharacteristic condition. Except for the integral on the boundary ($x = 0$), all of the integrals in

(3.12) are positive. The defect of this model is that there is no lower derivative in the double integral of (3.12), which prevents us from using energy estimates. However, this defect can be resolved by adding the following two conditions due to the nonlinearity of (1.1):

$$\text{for some constant } K_4,$$

$$0 < K_4(|\phi'| + |\psi'|) < -\mu(\phi,\psi)_x, \tag{3.13}$$

$$|\mu(\phi,\psi) - \sigma| \ll 1. \tag{3.14}$$

The reasoning behind these conditions can be found in Liu [9].

3.3. Step II: Preliminaries for the energy estimate. In the following, we assume (2.12) for $t = t_n$ a priori.

Integrate (3.8)z over $0 < \tau < t_n$ and $0 < x < \infty$.

$$\begin{aligned}
&\int_0^\infty \left(z_t z + (\lambda_1+\lambda_2)z_x z - \frac{h_v}{2}z^2\right)(x,t_n)dx + \int_0^t \left(-\lambda_1\lambda_2 z z_x - h_v\mu z^2\right)(0,\tau)d\tau \\
&+ \int_0^t\int_0^\infty -z_t^2 - \underbrace{((\lambda_1+\lambda_2)z)_t\, z_x}_{B_1} \underbrace{-(\lambda_1\lambda_2 z_x)z_x}_{B_2} dxd\tau + \int_0^t\int_0^\infty (h_{vt} + h_{vx}\mu)z^2 + h_v\mu_x z^2 dxd\tau \\
&\quad = \underbrace{\int_0^t\int_0^\infty f_v^{-1}\left(f_{vt} + g_v f_{vx} - g_{vx}\right)(z_t + f_u z_x)z dxd\tau}_{B_3} \\
&\qquad + \int_0^t\int_0^\infty (\text{RHS})'_{(3.8)} z dxd\tau + \int_0^\infty \left(z_t z + (\lambda_1+\lambda_2)z_x z - \frac{h_v z^2}{2}\right)(x,0)dx.
\end{aligned} \tag{3.15}$$

First, we evaluate B_3 as follows:

If δ_2 is sufficiently small, then from (3.13) and (3.14), we have

$$(h_{vt} + h_{vx}\mu)z^2 + h_v u_x z^2 \;= (h_{vv}\phi_x + h_{vu}\psi_x)(\mu - \sigma) + h_v\mu_x z^2 > \frac{1}{2}|h_v\mu_x|z^2. \tag{3.16}$$

Next, we evaluate B_1 by the following inequality. Since the magnitude of the shock is sufficiently small (i.e., $|u_- - u_+| \ll 1$), we have

$$|\phi'| + |\psi'| \gg |\phi''| + |\psi''|. \tag{3.17}$$

Hence

$$\begin{aligned}
B_1 &= \int_0^t\int_0^\infty ((\lambda_1+\lambda_2)z)_t z_x dxd\tau \\
&= \int_0^t\int_0^\infty (\lambda_1+\lambda_2)_t z z_x + (\lambda_1+\lambda_2)z_t z_x dxd\tau \\
&= O(1)(\|\phi'\|_\infty + \|\psi'\|_\infty)\int_0^t z^2(0,\tau)d\tau \\
&\qquad + \int_0^t\int_0^\infty (\lambda_1+\lambda_2)z_t z_x dxdz + o(1)\int_0^t\int_0^\infty (|\phi'| + |\psi'|)z^2 dxd\tau.
\end{aligned} \tag{3.18}$$

Similarly to (3.18), we have

$$\begin{aligned} B_2 &= \int_0^t \int_0^\infty -(\lambda_1\lambda_2 z)_x z_x dx d\tau \\ &= o(1)\int_0^t \int_0^\infty (|\phi'|+|\psi'|)z^2 dx d\tau \\ &\qquad + O(1)(\|\phi'\|_\infty + \|\psi'\|_\infty)\int_0^t z^2(0,\tau)d\tau - \int_0^t \int_0^\infty (\lambda_1\lambda_2)z_x^2 dx d\tau. \end{aligned} \tag{3.19}$$

Also, we have

$$\begin{aligned} B_3 &= \int_0^t \int_0^\infty f_v^{-1}\left(f_{vt}+g_v f_{vx} - g_{vx}\right)(z_t + f_u z_x) z dx d\tau \\ &= O(1)(\|\phi'\|_\infty + \|\psi'\|_\infty)\left(\int_0^\infty z^2(x,t)+z^2(x,0)dx + \int_0^t z^2(0,\tau)d\tau\right) \\ &\qquad + o(1)\int_0^\tau \int_0^\infty (|\phi'|+|\psi'|)z^2(x,\tau)dx d\tau. \end{aligned} \tag{3.20}$$

Using (3.16)–(3.20), we have that

$$\begin{aligned} &\int_0^\infty \left(z_t z + (\lambda_1+\lambda_2)z z_x - \frac{h_v}{2}z^2\right)(x,t)dx \\ &\quad + \int_0^t \int_0^\infty -z_t^2 - (\lambda_1+\lambda_2)z_t z_x - \lambda_1\lambda_2 z_x^2 + \frac{1}{2}|h_v\mu_x|z^2 dx d\tau \\ &\quad + \int_0^t (-\lambda_1\lambda_2 z z_x - h_v\mu z^2)(0,\tau)d\tau + \left(\frac{c_1+c_2}{2} - O(1)\delta_0\right)z^2(0,t) \\ &= \int_0^\infty \left(z_t z + (\lambda_1+\lambda_2)z_x z - \frac{h_v}{2}z^2\right)(x,0)dx \\ &\quad + O(1)\delta_2\bigg(\int_0^t \int_0^\infty (z_t^2 + z_x^2 + z_{tt}^2 + z_{xx}^2 + z_{xt}^2)dx d\tau \\ &\quad + \int_0^t z^2(0,\tau)d\tau + \int_0^t z^2(x,t)+z^2(x,0)dx\bigg) \\ &\quad + O(1)(\|\phi'\|_\infty + \|\psi'\|_\infty)\int_0^\infty z^2(0,\tau)d\tau. \end{aligned} \tag{3.21}$$

The double integrals on the RHS of (3.21) come from the nonlinear terms in $(\text{RHS})_{(3.8)}$. Note that the double integral in (3.21) has zero-order derivatives. Next, we use condition (3.10) and integrate $(3.8)z + (3.8)z_t + (3.8)((c_1+c_2)/2)z_x$ over the intervals $0 < x < \infty$ and $0 < \tau < t$. By the same method as used to derive (3.21), we have that

$$\begin{aligned} &(1-O(1)\delta_2)\bigg(\int_0^\infty \left(-\frac{\lambda_1\lambda_2}{2}z_x^2 + \frac{(z_t+(\lambda_1+\lambda_2)z_x)^2}{2} + \frac{(z_x+2z)^2}{2}\right)(x,t)dx \\ &\quad + \int_0^t \int_0^\infty \left(z_t^2 + 2\sigma z_x z_t + (\mu(\lambda_1+\lambda_2)-\lambda_1\lambda_2)z_x^2 + \frac{h_v\mu_x}{2}z^2\right)dx d\tau\bigg) \\ &\quad - (1+O(1)\delta_2)\frac{c_1+c_2}{2}z^2(0,t) + (1-O(1)\delta_2)\frac{c_1+c_2}{2}z^2(0,0) \end{aligned} \tag{3.22}$$

$$
\begin{aligned}
&+\int_0^t \left(\sigma(1+O(1)\delta_2)z^2-\lambda_1\lambda_2 z_x z\right)_{|x=0} d\tau\\
&+\int_0^t -\left(\frac{\lambda_1+\lambda_2}{2}z_x^2-\lambda_1\lambda_2 z_x z_t\left(-\frac{c_1+c_2}{2}-O(1)\delta_2\right)z_t^2\right)_{|x=0} d\tau\\
=&\,(1+O(1)\delta_2)\int_0^\infty \left(-\frac{\lambda_1\lambda_2}{2}z_x^2+\frac{(z_t+(\lambda_1+\lambda_2)z_x)^2}{2}+\frac{(z_x+2z)^2}{2}\right)(x,0)dx\\
&+O(1)\delta_2\int_0^t\int_0^\infty z_t^2+z_x^2+z_{xx}^2+z_{xt}^2+z_{tt}^2 dxd\tau.
\end{aligned}
$$

We denote

$$x_1 := x, \qquad x_2 := t.$$

Integrate $(3.22)_i$ and $(3.22)_{ij}$ below (for $i, j = 1, 2$) over $[0, \infty) \times [0, t_n]$. We have

(3.23)

$$\int_0^t\int_0^\infty (3.8)_{x_i} z_{x_i} + (3.8)_{x_i} z_{x_i x_2} + (3.8)_{x_i} z_{x_i x_1}\frac{(c_1+c_2)}{2} dxd\tau, \tag{$(3.23)_i$}$$

$$\int_0^t\int_0^\infty (3.8)_{x_i x_j} z_{x_i x_j} + (3.8)_{x_i x_j} z_{x_i x_j x_2} + (3.8)_{x_i x_j} z_{x_i x_j x_1}\frac{(c_1+c_2)}{2} dxd\tau. \tag{$(3.23)_{ij}$}$$

For the nonlinear terms on the RHS of $(3.23)_{ij}$, there are some integrals whose integrands have some terms of the fourth derivatives as follows:

$$\int_0^t\int_0^\infty W(z_{x_1}, z_{x_2}) z_{x_i x_j x_k x_l} z_{x_i x_j x_m} dxd\tau \qquad \text{for } i, j, m, l = 0,\ 1, \tag{3.24}$$

where $\lim_{(x^2+y^2)\to 0} W(x, y)/(x^2 + y^2)^{1/2}$ exists. If $m \in \{k, l\}$, then let $m = k$. Hence

$$
\begin{aligned}
&\int_0^t\int_0^\infty W(z_{x_1}, z_{x_2}) z_{x_i x_j x_k x_l} z_{x_i x_j x_k} dxd\tau\\
&\qquad = \frac{1}{2}\int_0^t\int_0^\infty W(z_{x_1}, z_{x_2})\partial_{x_l} z^2_{x_i x_j x_k} dxd\tau = (\mathrm{RHS})_{(3.25)}.
\end{aligned} \tag{3.25}
$$

By using integration by parts and the condition $|z_{x_p}| + |z_{x_i x_j}| \le N(t) + |u_- - u_+|$ for $p = 1, 2$,

$$
\begin{aligned}
(\mathrm{RHS})_{(3.25)} = O(1)\delta_2 \Bigg(&\int_0^t\int_0^\infty \left(z^2_{x_i x_j x_1} + z^2_{x_i x_j x_2}\right) dxd\tau + \int_0^t z^2_{x_i x_j x_k}(0, \tau)d\tau\\
&+\int_0^\infty z^2_{x_i x_j x_k}(x, t)dx + \int_0^\infty z^2_{x_i x_j x_k}(x, 0)dx\Bigg).
\end{aligned}
$$

If $m \ne k, l$, then we may assume that $m = 1$ and $k = l = 2$.

(3.26)

$$\begin{aligned}
&\int_0^t\int_0^\infty W(z_{x_1},z_{x_2})z_{x_ix_jx_2x_2}z_{x_ix_jx_1}dxd\tau\\
&=\int_0^t\int_0^\infty \big(\partial_{x_2}(W(z_{x_1},z_{x_2})z_{x_ix_jx_1}z_{x_ix_jx_2})-(\partial_{x_2}W(z_{x_1},z_{x_2}))z_{x_ix_jx_1}z_{x_ix_jx_2}\big)\,dxd\tau\\
&\qquad\qquad\qquad\qquad-\int_0^t\int_0^\infty W(z_{x_1},z_{x_2})z_{x_ix_jx_2x_1}z_{x_ix_jx_2}dxd\tau\\
&=(\mathrm{RHS})_{(3.26)}.
\end{aligned}$$

The last term on the RHS of (3.26) is handled similarly to (3.25). Consequently, the order of the other terms remaining on the RHS of (3.26) is at most three.

$(3.23)_{ij}$ can be rewritten as follows using (3.25) and (3.26):

$(3.23)'_{ij}$

$$\begin{aligned}
&(1-O(1)\delta_2)\bigg(\int_0^\infty -\frac{\lambda_1\lambda_2}{2}z^2_{x_ix_jx_1}+\frac{(z_{x_ix_jx_2}+(\lambda_1+\lambda_2)z_{x_ix_jx_1})^2}{2}(x,t)dx\\
&\qquad\qquad+\int_0^\infty\frac{(z_{x_ix_jx_1}+2z_{x_ix_j})^2}{2}(x,t)dx\\
&+\int_0^t\int_0^\infty z^2_{x_ix_jx_2}+2\sigma z_{x_ix_jx_1}z_{x_ix_jx_2}+(\mu(\lambda_1+\lambda_2)-\lambda_1\lambda_2)z^2_{x_ix_jx_1}+\frac{h_v\mu_x}{2}z^2_{x_ix_j}dxd\tau\bigg)\\
&\qquad-(1+O(1)\delta_2)\frac{c_1+c_2}{2}z^2_{x_ix_j}(0,t)-(1-O(1)\delta_2)\frac{c_1+c_2}{2}z^2_{x_ix_j}(0,0)\\
&\qquad+\int_0^t\Big(\sigma(1+O(1)\delta_2)z^2_{x_ix_j}-\lambda_1\lambda_2z_{x_ix_jx_1}z_{x_ix_j}\Big)(0,t)d\tau\\
&+\underbrace{\int_0^t-\bigg(\frac{\lambda_1+\lambda_2}{2}z^2_{x_ix_jx_1}-\lambda_1\lambda_2z_{x_ix_jx_1}z_{x_ix_jx_2}+\Big(-\frac{c_1+c_2}{2}-O(1)\delta_2\Big)z^2_{x_ix_jx_2}\bigg)(0,t)d\tau}_{I_{ij}}\\
&\qquad=(1+O(1)\delta_2)\int_0^\infty\bigg(-\frac{\lambda_1\lambda_2}{2}z^2_{x_ix_jx_1}+\frac{(z_{x_ix_jx_2}+(\lambda_1+\lambda_2)z_{x_ix_jx_1})^2}{2}\\
&\qquad\qquad+\frac{(z_{x_ix_jx_1}+2z_{x_ix_j})^2}{2}\bigg)(x,0)dx\\
&\qquad+O(1)\delta_2\int_0^t\int_0^\infty\sum_{i=1}^2z^2_{x_i}+\sum_{1\le i,j\le 2}z^2_{x_ix_j}+\sum_{1\le i,j,k\le 2}z^2_{x_ix_jx_k}dxd\tau.
\end{aligned}$$

In what follows, we look for suitable positive linear combinations of (3.22), $(3.23)_i$, and $(3.23)_{ij}$ to treat the integrals on the boundary $x=0$.

3.4. Step III: The estimate of the boundary integral at $x=0$. In (3.4) and (3.5), we have already estimated the lower derivative terms, i.e., those of order zero and one on the boundary $x=0$. From (3.8) and $z_x(0,t)=u_--\phi(-\sigma t-s_n)$, we have that

$$z_{tt}(0,t)=-(1+O(1)\delta_2)\lambda_1\lambda_2z_{xx}(0,t)+O(1)|\phi'(-\sigma t-s_n)|. \tag{3.27}$$

From (3.27), all of the second derivatives on the boundary $(x=0)$ can be expressed in terms of the linear combination of $z_{tt}(0,t)$ and the lower derivatives that have been estimated in (3.4) and (3.5). Similarly, for the third derivatives at the boundary $x = 0$, we have

$$z_{xtt}(0,t) = -\sigma^2 \phi''(-\sigma t - s_n). \tag{3.28}$$

From (3.28), $(3.8)_t$, and $(3.8)_x$, we have the following:

(3.29)

$$\begin{aligned} z_{ttt}(0,t) &= -(1 + O(1)\delta_2)(\lambda_1\lambda_2 z_{xxt}(0,t) + 2z_{tt}(0,t)) + O(1)\delta_2|z_t(0,t)|, \\ (\lambda_1+\lambda_2)z_{xt}(0,t) &= -(1+O(1)\delta_2)(\lambda_1\lambda_2 z_{xxx}(0,t) - 2\sigma z_{xx}(0,t)) + O(1)\delta_2|z_t(0,t)|. \end{aligned}$$

From (3.27) and (3.29), we can express the third derivatives $z_{x_i x_j x_k}(0,t)$ in terms of $z_{ttt}(0,t)$, $z_{tt}(0,t)$, and the other terms whose rates of convergence are known.

In $(3.23)'_{12}$, look at the integral that contains the third derivatives at the boundary $x = 0$ except for the quadratic term z_{xtt}:

(3.30)

$$\begin{aligned} &\int_0^t -\lambda_1\lambda_2 z_{xtt}(0,\tau) z_{xxt}(0,\tau) - \frac{\lambda_1+\lambda_2}{4}\lambda_1\lambda_2 z_{xxt}^2(0,\tau)d\tau \quad \left[\text{let } \epsilon(t) = \frac{\lambda_1+\lambda_2}{8} > 0\right] \\ &\quad \geq \int_0^t \frac{\lambda_1\lambda_2}{\epsilon(t)} z_{xtt}^2(0,\tau) + \epsilon(t)\lambda_1\lambda_2 z_{xxt}^2(0,\tau) - \frac{\lambda_1+\lambda_2}{4}\lambda_1\lambda_2 z_{xxt}^2(0,\tau)d\tau \\ &\quad \geq \int_0^t \frac{\lambda_1\lambda_2}{\epsilon(t)} z_{xtt}^2(0,\tau) - \frac{\lambda_1+\lambda_2}{8}\lambda_1\lambda_2 z_{xxt}^2(0,\tau)d\tau \quad [\text{from (3.28)}] \\ &\quad = O(1)\int_0^t z_t^2(0,\tau)d\tau - (1+O(1)\delta_2)\frac{c_1+c_2}{8(c_1+c_2)}\int_0^t \left(\frac{1}{2}z_{ttt}^2 - \frac{9}{c_1^2c_2^2}z_{tt}^2\right)d\tau. \end{aligned}$$

Conclusion. Using the inequalities in (3.30), we can find constants $D_{ij} > 0$ such that

$$|I_{ij}(t)| \leq D_{ij}\int_0^t \sum_{i=0}^3 (z_t)^2(0,\tau)d\tau, \tag{3.31}$$

where $I_{ij}(t)$ is the integral defined at the boundary $x = 0$ in $(3.23)'_{ij}$.

In order to evaluate the integral $\int_0^t z_t t^2(0,\tau)d\tau$, we need the following lemma.

Lemma 3.1. *The following inequality holds for an arbitrary positive constant C and an arbitrary function $q(t) \in C^3[0,\infty)$:*

$$\begin{aligned} &\sqrt{C}\int_0^t q_{tt}^2(0,\tau)d\tau + C^{1/4}q_{tt}^2(0,t) - 12(C^{3/4}+C^{1/4})\left(q_t^2(0,t) + q_{tt}^2(0,0) + q_{tt}^2(0,0)\right) \\ &\quad \leq C\int_0^t q_t^2(\tau)d\tau + 10\int_0^t q_{ttt}^2(\tau)d\tau. \end{aligned}$$

Proof. By using the Schwartz inequality and integration by parts,

(3.32)

$$\begin{aligned} &\int_0^t Cq_t^2 + q_{ttt}^2 d\tau \\ &\quad > \left|\int_0^t 2\sqrt{C}q_t q_{ttt}d\tau\right| \end{aligned}$$

$$
\begin{aligned}
&= \left| 2\sqrt{C}\int_0^t \partial_t(q_{tt}q_t)d\tau - 2\sqrt{C}\int_0^t q_{tt}^2 d\tau \right| \\
&\geq 2\sqrt{C}\left(\int_0^t q_{tt}^2 d\tau - |q_t q_{tt}(t)| - |q_t q_{tt}(0)| \right) \\
&\geq 2\sqrt{C}\left(\int_0^t q_{tt}^2 d\tau - (C^{1/4}q_t^2(t) + C^{-1/4}q_{tt}^2(t)) - \Big(C^{1/4}q_t^2(0) + C^{-1/4}q_{tt}^2(0) \Big) \right).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
(3.33) \qquad \int_0^t \sqrt{C}q_{tt}^2 + 9q_{ttt}^2 d\tau &> 6C^{1/4}\left| \int_0^t q_{tt}q_{ttt}d\tau \right| \\
&> 3C^{1/4}\left(q_{tt}^2(t) - q_{tt}^2(0) \right).
\end{aligned}
$$

The proof follows from (3.32) and (3.33).

From Lemma 3.1, we estimate the integral $\int_0^t z_{tt}(0,\tau)d\tau$ by $\int_0^t z_t^2(0,\tau)d\tau$ and $\int_0^t z_{ttt}^2(0,\tau)d\tau$. Namely, Lemma 3.1 allows us to bound $\int_0^t z_{tt}(0,\tau)d\tau$ by a positive linear combination of $\int_0^t z_t(0,\tau)d\tau$ and $\int_0^t z_{ttt}(0,\tau)d\tau$. Keeping the above calculation in mind, we complete the proof of Lemma 1.2. At first, we show that all the boundary terms can be treated precisely. To this end, we take the positive constant F_0 such that

$$
(3.34) \qquad -\frac{c_1+c_2}{8c_1c_2}F_0 > 2\sum_{1\leq i,j\leq 2} D_{ij}.
$$

Then from (3.34), (3.31), and (3.30), we have that

$$
\begin{aligned}
(3.35) \qquad &\sum_{(i,j)\neq(1,2),(2,1)} I_{ij}(t) + 2F_0 I_{12}(t) \\
&\qquad \geq -\frac{c_1+c_2}{16c_1c_2}\int_0^t z_{ttt}(0,\tau)d\tau - \sum_{(i,j)\neq(1,2),(2,1)} D_{ij}\sum_{i=0}^{2}\int_0^t (\partial_t^i z)^2(0,\tau)d\tau.
\end{aligned}
$$

From the positive linear combination in (3.35), we evaluate the integrals of the third derivatives at the boundary $x=0$. Actually, from Lemma 3.1, the integrals of the second derivatives at $x=0$ can be bounded by the third and first derivatives at the boundary. Hence for a large constant C,

$$
\begin{aligned}
(3.36) \qquad &C\int_0^t z_t^2(0,\tau)d\tau + \sum_{(i,j)\neq(1,2),(2,1)} I_{ij}(t) + 2F_0 I_{12}(t) \\
&\qquad \geq -O(1)C\int_0^t z^2(0,\tau)d\tau + \sqrt{\frac{C}{10}}\int_0^t z_{tt}^2(0,\tau)d\tau.
\end{aligned}
$$

Thus we finish the evaluation of the boundary integrals. Next, we complete the energy estimate for the following equation over the interval $(0,t]\times[0,\infty)$:

$$
(3.37) \qquad \sum_{(i,j)\neq(1,2),(2,1)} (3.23)'_{ij}z_{x_ix_j} + (3.23)'_{ij}z_{x_ix_jx_2} + \frac{c_1+c_2}{2}(3.23)'_{ij}z_{x_ix_jx_1}
$$

$$
\begin{aligned}
&+2F_0(3.23)'_{12}z_{x_1x_2}+2F_0(3.23)'_{12}z_{x_1x_2x_2}+2F_0\frac{c_1+c_2}{2}(3.23)'_{12}z_{x_1x_2x_1}\\
&+\sum_{1\le i\le 2}(3.23)_i z_{x_i}+(3.23)_i z_{x_ix_2}+\frac{c_1+c_2}{2}(3.23)_i z_{x_ix_1}\\
&+(3.23)z+(3.23)z_{x_2}+\frac{c_1+c_2}{2}(3.23)z_{x_1}.
\end{aligned}
$$

Hence by $(3.23)'_{ij}$, we have

$$
\begin{aligned}
&\int_0^\infty\left(\sum_{1\le i,j\le 2}z^2_{x_ix_j}+\sum_{1\le i\le 2}z_i^2+z^2\right)(x,t)dx+\int_0^t\int_0^\infty\sum_{i=1}^3|\nabla^i_{(x,t)}z|^2dxd\tau\\
&\quad\le O(1)\int_0^\infty\left(\sum_{1\le i,j\le 2}z^2_{x_ix_j}+\sum_{1\le i\le 2}z_i^2+z^2\right)(x,0)dx+O(1)C\int_0^t(z^2+z_t^2)(0,\tau)d\tau\\
&\quad=:(\mathrm{RHS})_{(3.38)}.
\end{aligned}
\tag{3.38}
$$

From (3.3), we have that

$$
\begin{cases}
|z(x,0)|=|(x_0-s_n)\phi'|+O(1)\delta_1|\phi(x-x_0)-u_+|/|u_--u_+| & \text{for } x>x_0,\\
|z(x,0)|=|(x_0-s_n)\phi'|+O(1)\delta_1x_0 & \text{for } 0<x<x_0.
\end{cases}
\tag{3.39}
$$

From (2.21), (3.4), and (3.5), if we choose δ_1, $e^{-|u_--u_+|x_0}$, and δ_1x_0 sufficiently small for fixed $\delta_2\ll 1$ and $|u_--u_+|\le\delta_2$, then by substituting (3.39) in (3.38), we get the a priori estimate

$$
N(t_n)+|u_--u_+|\le\frac{\delta_2}{2}.
\tag{3.40}
$$

At last, we have shown that (3.2) holds for n. This completes the proof of Lemma 1.2.

3.5. The convergence theorem (Theorem 1.4). Since we have the global existence theorem (Theorem 1.3) and (3.2) for all $n\in\mathbf{N}$, Theorem 1.4 is valid for $t=\infty$. This asserts the validity of asymptotic shock location $s_\infty+\sigma t$. Actually, replacing x_0 by s_m in (2.20), we have

$$
|s_n-s_m|\le\frac{K_4|\phi(-x_0-\sigma t_n)-\phi(-x_0-\sigma t_m)|}{|u_--u_+|^3}
$$

for some constant K_4. Hence $\{s_n\}_{n\in\mathbf{N}}$ is a Cauchy sequence, and thus $s_\infty=\lim_{n\to\infty}s_n$ exists. From the definition of $\{s_n\}_{n\in\mathbf{N}}$, we have that

$$
\lim_{t\to\infty}\int_0^\infty(u(x,0)-\phi(x-\sigma t-s_\infty))\,dx=0.
$$

Let $z_\infty(x,t):=-\int_x^\infty u(x,t)-\phi(x-\sigma t-s_\infty)dx$ and $w_\infty(x,t):=v(x,t)-\psi(x-\sigma t-s_\infty)$. From assumption (1.8) of the initial value, we obtain

$$
\int_0^\infty\sum_{j=1}^3|\partial_x^jz_\infty|^2(x,0)dx+\int_0^\infty\sum_{j=0}^2|\partial_x^jw_\infty|^2(x,0)dx
$$

arbitrarily small when we choose δ_1 and x_0 to have $\exp(-x_0|u_- - u_+|)$, where $\delta_1 x_0$ and δ_1 are sufficiently small. By the same method as used to show global existence, we have the following energy estimate:

$$\int_0^\infty \int_0^\infty \sum_{j=1}^{3} \left(|\partial_x^j z_\infty|^2 + |\partial_t^j z_\infty|^2\right)(x,\tau)dxd\tau + \int_0^\infty z_\infty^2(x,t)dx < C, \tag{3.41}$$

where C is a constant independent of t. By making use of (3.41), it follows that the solution converges to the asymptotic traveling wave $(\phi(x-\sigma t - s_\infty), \psi(x - \sigma t - s_\infty))$.

Acknowledgment. The authors would like to express their gratitude to their thesis advisor, Professor Tai-Ping Liu, for generously sharing his ideas.

REFERENCES

[1] R. E. Caflish, *The fluid dynamics limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math., 33 (1980), pp. 651–666.

[2] R. E. Caflish and G. C. Papanicolaou, *The fluid dynamics limit of the nonlinear model Boltzmann equation*, Comm. Pure Appl. Math., 32 (1979), pp. 589–616.

[3] G.-Q. Chen and T.-P. Liu, *Zero relaxation and dissipation limits for hyperbolic conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 755–781.

[4] G.-Q. Chen, C. D. Levemore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[5] S. Kawashima and A. Matsumura, *Asymptotic stability of traveling wave solution of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.

[6] T.-T. Li and W.-C. Yu, *Boundary value problems for the first order quasilinear hyperbolic systems and their applications*, J. Differential Equations, 41 (1981), pp. 1–26.

[7] T.-T. Li and W.-C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Math. Ser. V, Duke University Press, Durham, NC.

[8] J. G. Liu and Z. Xin, *Boundary layer behavior in the fluid-dynamic limit for a nonlinear model Boltzmann equation*, Arch. Rational Mech. Anal., to appear.

[9] T.-P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[10] T.-P. Liu and S. Yu, *The propagation of the stationary viscous shock under the boundary effect*, to appear.

[11] T. P. Liu and K. Zumbrun, *Nonlinear stability of an undercompressible shock for the complex Burgers equation*, to appear.

[12] S. Nishibata, *The initial boundary problem for hyperbolic conservation laws with relaxation*, J. Differential Equations, 130 (1996), pp. 100–126.

[13] Z. Xin, *The fluid-dynamics limit of the Browdwell model of the nonlinear Boltzmann equation in the presence of shock*, Comm. Pure Appl. Math., 44 (1991), pp. 679–713.

[14] S. Yu, *The asymptotic behavior of the Burgers equation on the quarter plane*, to appear.

[15] G. B. Whitham, *The Linear and Nonlinear Waves*, John Wiley, New York, 1974.

[16] I.-L. Chen, *Long-time effect of relaxation for the hyperbolic conservation laws*, to appear.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 322–337, March 1997

A NONLINEAR GRATING PROBLEM IN DIFFRACTIVE OPTICS[*]

GANG BAO[†] AND YUNMEI CHEN[†]

Abstract. In this paper, the existence and regularity of solutions of a system of Maxwell equations in periodic structures are established. The diffraction (scattering) problem is reformulated in a bounded domain by introducing the transparent boundary conditions. An indirect approach is then used to carry out the regularity analysis. This problem arises in the modeling of the surface-enhanced second harmonic generation of nonlinear optics.

Key words. existence and regularity of solutions, Maxwell equations, second harmonic generation, nonlinear grating

AMS subject classifications. 35J25, 78A45, 78A60

PII. S0036141095284461

1. Introduction. Consider a plane wave of frequency ω_1 incident on a grating or periodic structure ruled on some nonlinear optical material. Because of the presence of the nonlinear material, the nonlinear optical interaction gives rise to diffracted waves of frequencies ω_1 and $\omega_2 = 2\omega_1$. This process represents the simplest situation in nonlinear optics—second harmonic generation (SHG). In this paper, we study questions on the existence and regularity of solutions of the PDE that governs SHG. As a first step, we use the well-known undepleted-pump approximation, which allows us to reduce the model to a system of generalized Helmholtz equations. We are seeking a "quasi-periodic" solution at frequency ω_1 or ω_2 which satisfies the "radiation condition at infinity"; this means that near $x_2 = \pm\infty$, the solution is a linear combination of plane waves propagating away from the structure.

This work is motivated by the recent research on surface-enhanced nonlinear optical effects. One of the many important applications of nonlinear optical phenomena is a method for obtaining coherent radiation at a wavelength shorter than that of available lasers through SHG. Unfortunately, nonlinear optical effects are generally so weak that their observation requires extremely high-intensity laser beams. Recently, in the sequence of papers [13], [14], [12], a PDE model was introduced to model nonlinear SHG in periodic structures. In particular, it was shown in [13] and [14] that SHG can be greatly enhanced by using diffraction gratings or periodic structures and that the PDE model can accurately predict the field propagation. Our goal is to examine the well-posedness of the PDE model by using our variational formulation.

Our main well-posedness result is as follows: there exists a unique quasi-periodic solution of frequencies ω_1 and $\omega_2 = 2\omega_1$ of the nonlinear diffraction problem with the radiation condition for all but possibly a discrete set of parameters. The proof of this result follows in principle an earlier work of DiBenedetto, Elliott, and Friedman [6]. However, because of weaker regularity assumptions on the coefficients in our model PDE, essential modifications must be made.

Results on existence and uniqueness for Maxwell equations in linear media with

[*] Received by the editors April 10, 1995; accepted for publication (in revised form) November 29, 1995.

http://www.siam.org/journals/sima/28-2/28446.html

[†] Department of Mathematics, University of Florida, Gainesville, FL 32611 (bao@math.ufl.edu, yun@math.ufl.edu). The research of the first author was partially supported by NSF grant DMS-9501099 and a University of Florida Research Development Award. The research of the second author was partially supported by NSF grant DMS-9123532.

periodic structures were obtained by Chen and Friedman [5] assuming that the dielectric coefficient ϵ is a piecewise-constant function. They showed that for all but possibly a discrete number of ϵ's, there exists a unique solution to the Maxwell equations by an integral-equation approach. Little is known concerning the questions of existence and uniqueness for nonlinear Maxwell equations in periodic structures. In two simple cases where the Maxwell equations can be reduced to a system of nonlinear Helmholtz equations, existence and uniqueness results have recently been obtained in [2] and [3]. Computational results have also been obtained by using a combination of the method of finite elements and the fixed-point iteration algorithm. In this work, using a different approach, we study the well-posedness of a more complicated but linear PDE model. Its main advantage over previous results is that the current model supports a larger class of nonlinear optical materials with cubic symmetry structures. An interesting future project is to tackle the nonlinear model directly.

A good background on the linear theory of diffractive optics in grating structures may be found in Petit [11]. A brief description of the present problem along with a discussion of some other mathematical problems arising from industrial applications of diffractive optics can be found in Friedman [7, Chapter 5]. For the underlying physics of nonlinear optics, we refer the reader to the classic books of Bloembergen [4] and Shen [15].

The outline of this paper is as follows. The governing system of Maxwell equations is introduced in section 2, and it is reformulated inside a periodic "box" with boundary conditions derived from our knowledge of the fundamental solutions in linear media. We then proceed in section 3 to study the corresponding PDEs of frequencies ω_1 and $\omega_2 = 2\omega_1$ by a variational approach. The existence and regularity of solutions of the diffraction problem are established.

2. Modeling of the scattering problem. Throughout, the media are assumed to be nonmagnetic with constant magnetic permeability. For convenience, the magnetic permeability constant is assumed to be equal to unity everywhere. We also assume that no external charge or current is present in the field.

The time-harmonic Maxwell equations that govern SHG then take the form

$$\nabla \times \mathbf{E} = -\frac{i\omega}{c}\mathbf{H}, \qquad \nabla \cdot \mathbf{H} = 0, \tag{2.1}$$

$$\nabla \times \mathbf{H} = \frac{i\omega}{c}\mathbf{D}, \qquad \nabla \cdot \mathbf{D} = 0, \tag{2.2}$$

along with the constitutive equation

$$\mathbf{D} = \epsilon\mathbf{E} + 4\pi\chi^{(2)}(x,\omega) : \mathbf{E}\mathbf{E}, \tag{2.3}$$

where $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, $\mathbf{D}$ is electric displacement, ϵ is the dielectric coefficient, c is the speed of light, ω is the angular frequency, and $\chi^{(2)}$ is the second-order nonlinear susceptibility tensor of third rank, i.e., $\chi^{(2)} : \mathbf{E}\mathbf{E}$ is a vector whose jth component is $\sum_{k,l=1}^{3} \chi^{(2)}_{jkl}\mathbf{E}_k\mathbf{E}_l$.

Remarks. The medium is said to be linear if $\mathbf{D} = \epsilon\mathbf{E}$ or $\chi^{(2)}$ vanishes. In principle, all optical media are essentially nonlinear, i.e., $\mathbf{D}$ is a nonlinear function of $\mathbf{E}$.

The physics of SHG may be described as follows: when a plane wave at frequency $\omega = \omega_1$ is incident on a nonlinear medium, because of the interaction of the incident wave and nonlinear medium, diffracted waves of frequencies $\omega = \omega_1$ and $\omega = 2\omega_1$ are generated. The fact that new frequency components are present is the most striking difference between nonlinear and linear optics. However, for most media, nonlinear optical effects are so weak that they may reasonably be ignored. In particular, the conversion of energy into the new frequency component is very small. The observation of nonlinear phenomena in the optical region can normally only be made by using high-intensity beams, say by application of a high-intensity laser.

We assume that the depletion of energy from the pump waves (at frequency $\omega = \omega_1$) may be neglected, which is the well-known undepleted-pump approximation in the literature; see [13] and [14]. Under the approximation, equation (2.3) at frequencies $\omega = \omega_1$ and $\omega = \omega_2 = 2\omega_1$, respectively, may be written as

$$\mathbf{D}(x,\omega_1) = \epsilon(x,\omega_1)\mathbf{E}(x,\omega_1), \tag{2.4}$$

$$\mathbf{D}(x,\omega_2) = \epsilon(x,\omega_2)\mathbf{E}(x,\omega_2) + 4\pi\chi^{(2)}(x,\omega_2) : \mathbf{E}(x,\omega_1)\mathbf{E}(x,\omega_1). \tag{2.5}$$

Next, we want to further reduce the nonlinear coupled system (2.1)–(2.2). Throughout the paper, all of the fields are assumed to be invariant in the x_3 direction. Here, as in the linear case, in transverse electric (TE) polarization the electric field is transversal to the (x_1, x_2)-plane and in transverse magnetic (TM) polarization the magnetic field is transversal to the (x_1, x_2)-plane. In the nonlinear case, however, the polarization is determined by group symmetry properties of $\chi^{(2)}$. In this paper, motivated by applications, we assume that the electromagnetic fields are TM polarized at frequency ω_1 and TE polarized at frequency ω_2. This polarization assumption is known to support a large class of nonlinear optical materials, for example, crystals with cubic symmetry structures.

Therefore,

$$\mathbf{H}(x,\omega_1) = u(x_1,x_2,\omega_1)\vec{x_3}, \tag{2.6}$$

$$\mathbf{E}(x,\omega_2) = v(x_1,x_2,\omega_2)\vec{x_3}. \tag{2.7}$$

For convenience, we define

$$\epsilon_j = \epsilon(x_1,x_2,\omega_j), \quad j = 1,2, \tag{2.8}$$

$$k_j = \frac{\omega_j}{c}\sqrt{\epsilon_j}, \quad \mathrm{Im} k_j \geq 0. \tag{2.9}$$

At frequency ω_1, system (2.1)–(2.2) can be simplified to

$$\nabla \cdot \left(\frac{1}{k_1^2} \nabla u \right) + u = 0. \tag{2.10}$$

Because of equation (2.2),

$$\begin{aligned} \mathbf{E}(x,\omega_1) &= \frac{c}{i\omega_1\epsilon_1} \nabla \times \mathbf{H}(x,\omega_1) \\ &= \frac{c}{i\omega_1\epsilon_1}(\partial_{x_2}u, -\partial_{x_1}u, 0). \end{aligned}$$

Hence the second harmonic field satisfies

$$\left[\triangle + k_2^2\right] v = -\frac{4\pi\omega_2^2}{c^2} \sum_{j,l=1,2,3} \chi^{(2)}_{3jl}(x,\omega_2)(\mathbf{E}(x,\omega_1))_j(\mathbf{E}(x,\omega_1))_l \tag{2.11}$$

$$= \sum_{j,l=1,2} \chi_{jl}\partial_{x_j}u\,\partial_{x_l}u, \tag{2.12}$$

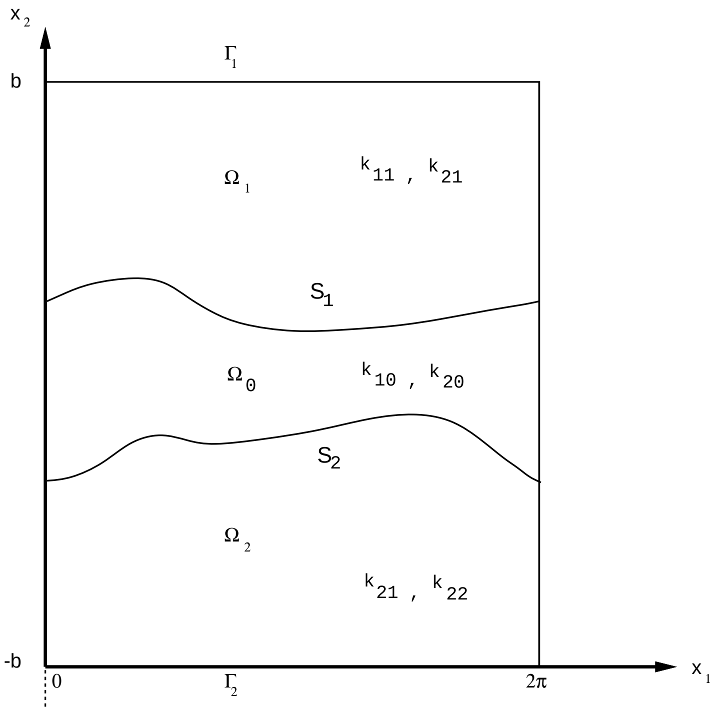


FIG. 1. *Problem geometry.*

where $\triangle$ is the usual Laplace operator and $\chi_{jl} = (-1)^{j+l}(16\pi/\epsilon_1^2)\chi^{(2)}_{3,j,l}(x,\omega_2)$.

Let us further specify the problem geometry. We assume that the medium and material are periodic in the x_1 variable of period 2π and are invariant in the x_3 variable. We may then restrict ourselves to one period in x_1, as shown in Figure 1.

We introduce the following notation:

$$\begin{aligned} \Gamma_j &= \{x_2 = (-1)^{j-1}b,\ 0 < x_1 < 2\pi\}, \qquad S_j = \{0 < x_1 < 2\pi,\ x_2 = \phi_j(x_1)\},\\ \Omega_1 &= \{0 < x_1 < 2\pi,\ \phi_1(x_1) < x_2 < b\}, \qquad \Omega_2 = \{0 < x_1 < 2\pi,\ -b < x_2 < \phi_2(x_1)\},\\ \Omega_1^+ &= \{0 < x_1 < 2\pi,\ x_2 \geq b\}, \qquad \Omega_2^+ = \{0 < x_1 < 2\pi,\ x_2 \leq -b\},\\ \Omega_0 &= \{0 < x_1 < 2\pi,\ \phi_2(x_1) < x_2 < \phi_1(x_1)\}, \qquad \Omega = \{0 < x_1 < 2\pi,\ -b < x_2 < b\}. \end{aligned}$$

Suppose that the whole space is filled with material in such a way that the “indexes of refraction” k_1 and k_2 satisfy

$$k_j(x) = \begin{cases} k_{j1} & \text{in } \Omega_1^+ \cup \bar{\Omega}_1,\\ k_{j0} & \text{in } \Omega_0,\\ k_{j2} & \text{in } \Omega_2^+ \cup \bar{\Omega}_2, \end{cases}$$

for $j = 1, 2$, where k_{j1} and k_{j2} are constants, k_{j1} are real and positive, and $\text{Re}k_{j2} > 0$ and $\text{Im}k_{j2} \geq 0$. The case $\text{Im}k_{j2} > 0$ accounts for materials which absorb energy.

Throughout, we make the following regularity assumptions on the material and geometry:

$$(2.13) \qquad k_{j0}(x) \in C^1(\Omega_0), \qquad \phi_j(x_1) \in C^{1,\gamma}(0, 2\pi) \quad \text{for some } 0 < \gamma < 1,$$

$$\chi_j \in L^\infty(\Omega) \quad \text{and} \quad \chi_j = 0 \quad \text{in } \Omega_1 \cup \Omega_2. \tag{2.14}$$

We wish to solve system (2.10)–(2.12) when an incoming plane wave

$$u_I = u_i e^{i\alpha_1 x_1 - i\beta_{11} x_2} \tag{2.15}$$

is incident on S_1 from Ω_1^+, where u_i is a real constant, $\alpha_1 = k_{11}\sin\theta$, $\beta_{11} = k_{11}\cos\theta$, and $-\pi/2 < \theta < \pi/2$ is the angle of incidence.

We are interested in "quasi-periodic" solutions (u, v), that is, solutions (u, v) such that

$$u_\alpha = u e^{-i\alpha_1 x_1} \quad \text{and} \quad v_\alpha = v e^{-i\alpha_2 x_1} \quad (\alpha_2 = k_{21}\sin\theta)$$

are 2π-periodic in the x_1 direction.

It follows from system (2.10)–(2.12) that

$$\nabla_{\alpha_1} \cdot \left(\frac{1}{k_1^2} \nabla_{\alpha_1} u_\alpha\right) + u_\alpha = 0, \tag{2.16}$$

$$\left[\triangle_{\alpha_2} + k_2^2\right] v_\alpha = \sum_{j,l=1,2} \chi_{jl}^\alpha \partial_{x_j}^{\alpha_1} u \, \partial_{x_l}^{\alpha_1} u, \tag{2.17}$$

where

$$\triangle_{\alpha_2} = \triangle + 2i\alpha_2 \partial_{x_1} - |\alpha_2|^2, \qquad \nabla_{\alpha_1} = \nabla + i(\alpha_1, 0),$$

and

$$\chi_{jl}^\alpha = \chi_{jl} e^{i(2\alpha_1 - \alpha_2)x_1}, \qquad \partial_{x_1}^{\alpha_1} = \partial_{x_1} + i\alpha_1, \qquad \partial_{x_2}^{\alpha_1} = \partial_{x_2}.$$

We next derive the transparent boundary conditions.

Expand u_α and v_α in a Fourier series:

$$u_\alpha(x_1, x_2) = \sum_{n \in Z} u_\alpha^n(x_2) e^{inx_1}, \tag{2.18}$$

$$v_\alpha(x_1, x_2) = \sum_{n \in Z} v_\alpha^n(x_2) e^{inx_1}, \tag{2.19}$$

where

$$u_\alpha^n(x_2) = \frac{1}{2\pi} \int_0^{2\pi} u_\alpha(x_1, x_2) e^{-inx_1} dx_1,$$
$$v_\alpha^n(x_2) = \frac{1}{2\pi} \int_0^{2\pi} v_\alpha(x_1, x_2) e^{-inx_1} dx_1.$$

For $j = 1, 2$, define the coefficients

$$\beta_{1j}^n(\alpha) = e^{i\gamma_{1j}/2} \left|k_{1j}^2 - (n + \alpha_1)^2\right|^{1/2}, \quad n \in Z, \tag{2.20}$$

$$\beta_{2j}^n(\alpha) = e^{i\gamma_{2j}/2} \left|k_{2j}^2 - (n + \alpha_2)^2\right|^{1/2}, \quad n \in Z, \tag{2.21}$$

where

$$\gamma_{1j} = \arg(k_{1j}^2 - (n + \alpha_1)^2), \quad 0 \le \gamma_{1j} < 2\pi$$
$$\gamma_{2j} = \arg(k_{2j}^2 - (n + \alpha_2)^2), \quad 0 \le \gamma_{2j} < 2\pi.$$

Throughout, we assume that $k_{1j}^2 \neq (n+\alpha_1)^2$ and $k_{2j}^2 \neq (n+\alpha_2)^2$ for all $n \in Z$, $j = 1, 2$; this assumption excludes the "resonant" cases, where waves can propagate along the x_1-axis. The assumption also ensures that a fundamental solution for (2.16) and (2.17) exists inside Ω_1 and Ω_2. It then follows from our knowledge of the fundamental solution (see, e.g., [5]) that inside Ω_1 and Ω_2, u_α and v_α can be expressed as a sum of plane waves: for $j = 1, 2$,

$$u_\alpha|_{\Omega_j} = \sum_{n\in Z} a_j^n e^{\pm i\beta_{1j}^n(\alpha)x_2 + inx_1}, \tag{2.22}$$

$$v_\alpha|_{\Omega_j} = \sum_{n\in Z} b_j^n e^{\pm i\beta_{2j}^n(\alpha)x_2 + inx_1}, \tag{2.23}$$

where a_j^n and b_j^n are complex scalars. Since β_{sj}^n is real for at most finitely many n's, there are only a finite number of *propagating* plane waves in the sum (2.22) and (2.23); the remaining waves are exponentially damped (or unbounded) as $|x_2| \to \infty$. We are only interested in the case where (u_α, v_α) is composed of bounded *outgoing* plane waves in Ω_1 and Ω_2 plus the incident incoming wave u_I in Ω_1. From (2.18), (2.19), (2.22), and (2.23), we have

$$u_\alpha^n(x_2) = \begin{cases} u_\alpha^n(b)e^{i\beta_{11}^n(\alpha)(x_2-b)}, & n \neq 0, \quad \text{in } \Omega_1, \\ u_\alpha^0(b)e^{i\beta_{11}(x_2-b)} + u_i e^{-i\beta_{11}x_2} - u_i e^{i\beta_{11}(x_2-2b)}, & n = 0, \quad \text{in } \Omega_1, \\ u_\alpha^n(-b)e^{-i\beta_{12}^n(\alpha)(x_2+b)} & \qquad\quad\;\; \text{in } \Omega_2 \end{cases} \tag{2.24}$$

and

$$v_\alpha^n(x_2) = \begin{cases} v_\alpha^n(b)e^{i\beta_{21}^n(\alpha)(x_2-b)} & \text{in } \Omega_1, \\ v_\alpha^n(-b)e^{-i\beta_{22}^n(\alpha)(x_2+b)} & \text{in } \Omega_2. \end{cases} \tag{2.25}$$

Let ν be the unit outward normal vector defined by

$$\nu = \begin{cases} \vec{x_2} & \text{on } \Gamma_1, \\ -\vec{x_2} & \text{on } \Gamma_2, \end{cases}$$

where $\vec{x_2}$ is the unit vector of the x_2-axis.

We can then calculate the derivative of $u_\alpha^n(x_2)$ and $v_\alpha^n(x_2)$ with respect to ν:

$$\left.\frac{\partial u_\alpha^n}{\partial \nu}\right|_{\Gamma_j} = \begin{cases} i\beta_{11}^n u_\alpha^n(b), & n \neq 0, \quad \text{on } \Gamma_1, \\ i\beta_{11} u_\alpha^0(b) - 2iu_i\beta_{11}e^{-i\beta_{11}b}, & n = 0, \quad \text{on } \Gamma_1, \\ i\beta_{12}^n u_\alpha^n(-b) & \qquad\quad\;\; \text{on } \Gamma_2 \end{cases} \tag{2.26}$$

and

$$\left.\frac{\partial v_\alpha^n}{\partial \nu}\right|_{\Gamma_j} = \begin{cases} i\beta_{21}^n(\alpha)v_\alpha^n(b) & \text{on } \Gamma_1, \\ i\beta_{22}^n(\alpha)v_\alpha^n(-b) & \text{on } \Gamma_2. \end{cases} \tag{2.27}$$

Therefore,

$$\left.\frac{\partial u_\alpha}{\partial \nu}\right|_{\Gamma_1} = \sum_{n\in Z} i\beta_{11}^n u_\alpha^n(b)e^{inx_1} - 2iu_i\beta_{11}e^{-i\beta_{11}b}, \tag{2.28}$$

$$\left.\frac{\partial u_\alpha}{\partial \nu}\right|_{\Gamma_2} = \sum_{n\in Z} i\beta_{12}^n u_\alpha^n(-b)e^{inx_1}, \tag{2.29}$$

$$\left.\frac{\partial v_\alpha}{\partial \nu}\right|_{\Gamma_1} = \sum_{n\in Z} i\beta_{21}^n v_\alpha^n(b)e^{inx_1}, \tag{2.30}$$

$$\left.\frac{\partial v_\alpha}{\partial \nu}\right|_{\Gamma_2} = \sum_{n\in Z} i\beta_{22}^n v_\alpha^n(-b)e^{inx_1}. \tag{2.31}$$

Since the fields u_α and v_α are 2π-periodic in x_1, without loss of generality, we can move the problem from $\mathbb{R}^2$ to the quotient space (cylinder) $\mathbb{R}^2/(2\pi Z \times \{0\})$. For the remainder of the paper, we shall identify Ω with the cylinder $\Omega/(2\pi Z \times \{0\})$ and do similarly for the boundaries $\Gamma_j \equiv \Gamma_j/2\pi Z$. Thus from now on, all functions defined on Ω and Γ_j are implicitly 2π-periodic in the x_1 variable.

For functions $f \in H^{1/2}(\Gamma_j)$ (the Sobolev space of complex-valued functions on the circle), define the operator T_{sj}^α by

$$(T_{sj}^\alpha f)(x_1) = \sum_{n\in Z} i\beta_{sj}^n(\alpha) f^n e^{inx_1} \tag{2.32}$$

for $s, j = 1, 2$, where $f^n = (1/2\pi)\int_0^{2\pi} f(x_1)e^{-inx_1}$ and equality is taken in the sense of distributions.

From (2.32) and the definition of $\beta_{sj}^n(\alpha)$, it is clear that T_{sj}^α is a standard pseudodifferential operator (in fact, a convolution operator) of order one.

From (2.28)–(2.31), we see that for $j = 1, 2$,

$$\begin{aligned} T_{11}^\alpha(u_\alpha|_{\Gamma_1}) &= \frac{\partial u_\alpha}{\partial \nu}|_{\Gamma_1} + 2iu_i\beta_{11}e^{-i\beta_{11}b}, \\ T_{12}^\alpha(u_\alpha|_{\Gamma_2}) &= \frac{\partial u_\alpha}{\partial \nu}|_{\Gamma_2}, \\ T_{2j}^\alpha(v_\alpha|_{\Gamma_j}) &= \frac{\partial v_\alpha}{\partial \nu}|_{\Gamma_j}, \end{aligned}$$

that is, T_{sj}^α are Dirichlet–Neumann maps. We will use the abbreviated notations $T_{1j}^\alpha u_\alpha$ and $T_{2j}^\alpha v_\alpha$ to mean $T_{1j}^\alpha(u_\alpha|_{\Gamma_j})$ and $T_{2j}^\alpha(v_\alpha|_{\Gamma_j})$, respectively.

Thus the scattering problem has been formulated as follows:

$$\nabla_{\alpha_1} \cdot \left(\frac{1}{k_1^2} \nabla_{\alpha_1} u_\alpha \right) + u_\alpha = 0 \quad \text{in } \Omega, \tag{2.33}$$

$$(\triangle_{\alpha_2} + k_2^2)v_\alpha = \sum_{j,l=1,2} \chi_{jl}^\alpha \partial_{x_j}^{\alpha_1} u_\alpha \partial_{x_l}^{\alpha_1} u_\alpha \quad \text{in } \Omega, \tag{2.34}$$

$$\left(T_{11}^\alpha - \frac{\partial}{\partial \nu} \right) u_\alpha = 2iu_i\beta_{11}e^{-i\beta_{11}b} \quad \text{on } \Gamma_1, \tag{2.35}$$

$$\left(T_{12}^\alpha - \frac{\partial}{\partial \nu} \right) u_\alpha = 0 \quad \text{on } \Gamma_2, \tag{2.36}$$

$$\left(T_{21}^\alpha - \frac{\partial}{\partial \nu} \right) v_\alpha = 0 \quad \text{on } \Gamma_1, \tag{2.37}$$

$$\left(T_{22}^\alpha - \frac{\partial}{\partial \nu} \right) v_\alpha = 0 \quad \text{on } \Gamma_2. \tag{2.38}$$

3. Existence and regularity of the scattering problem.

We are now ready to present the main result of this paper.

THEOREM 3.1. *Assume that the regularity assumptions* (2.13) *and* (2.14) *hold. Then for all but possibly a discrete set of frequencies ω, the scattering problem* (2.33)–(2.38) *admits a unique solution*

$$u_\alpha \in H^1(\Omega) \cap C^{1,\sigma}(\Omega_1^+ \cup \Gamma_1) \cap C^{1,\sigma}(\overline{\Omega}_0) \cap C^{1,\sigma}(\Omega_2^+ \cup \Gamma_2),$$

and $v_\alpha \in W^{2,p}(\Omega)$ for some $0 < \sigma < 1$ and any $1 < p < \infty$.

The proof may be given by proving Theorems 3.3 and 3.10 below.

We first establish $C^{1,\sigma}$ estimates of the pump field u_α. As in [1], we introduce a smooth 2π-periodic function u_0 supported near Γ_1 such that $u_0(x_1, b) = 0$ and $\partial_{x_2} u_0(x_1, b) = -2iu_i\beta_{11}e^{-i\beta_{11}b}$. Recall that u_i is the fixed constant given in (2.15). Then $u_\alpha - u_0$, still denoted by u_α, satisfies

$$\nabla_{\alpha_1} \cdot \left(\frac{1}{k_1^2} \nabla_{\alpha_1} u_\alpha\right) + u_\alpha = -f \quad \text{in } \Omega, \tag{3.1}$$

$$\left(T_{11}^\alpha - \frac{\partial}{\partial \nu}\right) u_\alpha = 0 \quad \text{on } \Gamma_1, \tag{3.2}$$

$$\left(T_{12}^\alpha - \frac{\partial}{\partial \nu}\right) u_\alpha = 0 \quad \text{on } \Gamma_2, \tag{3.3}$$

where

$$f = \nabla_{\alpha_1} \cdot \left(\frac{1}{k_1^2} \nabla_{\alpha_1} u_0\right).$$

By examining the variational formulation of (3.1)–(3.3), the following result holds; see [1] for the proof.

THEOREM 3.2. *Under the assumptions* (2.13) *and* (2.14)*, for all but possibly a discrete set of frequencies ω, the problem* (3.1)–(3.3) *admits a unique weak solution $u_\alpha \in H^1(\Omega)$.*

According to the standard elliptic regularity theory [9], [10], $u_\alpha \in C^\beta(\tilde{\Omega})$ for some $0 < \beta < 1$ and $||u_\alpha||_{C^\beta(\tilde{\Omega})} \le C$ with $\tilde{\Omega} \Subset \Omega$, where C depends on k_1, f, and $\text{dist}(\tilde{\Omega}, \Omega)$.

Denote $\tilde{\Omega} = \{(x_1, x_2),\ 0 < x_1 < 2\pi,\ -b_1 < x_2 < b_1\}$, where $b_1 < b$ is chosen such that $\tilde{\Omega} \subseteq \Omega \backslash \text{supp}\{u_0\}$ and $-b_1 < \phi_2(x) < \phi_1(x) < b_1\ \forall x_1 \in (0, 2\pi)$.

THEOREM 3.3. *Under the assumptions* (2.13) *and* (2.14)*, for all but possibly a discrete set of frequencies ω, the problem* (3.1)–(3.3) *admits a unique solution*

$$u_\alpha \in H^1(\Omega) \cap C^{1,\sigma}(\Omega_1^+ \cup \Gamma_1) \cap C^{1,\sigma}(\overline{\Omega}_0) \cap C^{1,\sigma}(\Omega_2^+ \cup \Gamma_2)$$

for some $0 < \sigma < 1$.

Moreover,

$$\|\nabla u_\alpha\|_{C^\sigma(\Omega_1^+ \cup \Gamma_1)} + \|\nabla u_\alpha\|_{C^\sigma(\overline{\Omega}_0)} + \|\nabla u_\alpha\|_{C^{1,\sigma}(\Omega_2^+ \cup \Gamma_2)} \le C, \tag{3.4}$$

where the constant C depends on k_1, α_1, u_i, and $\|u_\alpha\|_{C^\gamma(\tilde{\Omega})}$.

Note that near the boundary Γ_1 or Γ_2, the problem (3.1)–(3.3) becomes a Helmholtz equation with constant coefficients and smooth pseudodifferential boundary conditions. The $C^{1,\sigma}$ regularity of u_α then follows.

The standard elliptic regularity results [9] indicate the $C^{1,\sigma}$ regularity of u_α away from a tubular neighborhood of the two interfaces S_1 and S_2 due to the definition of k_1 and (2.13).

Thus it suffices to prove the regularity result in the tubular neighborhood of each interface. We shall restrict our attention to the interface S_1. The $C^{1,\sigma}$ regularity of u_α in the tubular neighborhood of S_2 can be established in the same manner; hence its proof will be omitted.

For simplicity, we shall drop the subscript of u_α. By using the definition of ∇_{α_1} and $f = 0$ away from near S_1, equation (3.1) may be rewritten as

$$\nabla \cdot \left(\frac{1}{k_1^2} \nabla u\right) + i\alpha_1 \frac{1}{k_1^2} \frac{\partial u}{\partial x_1} + i\alpha_1 \frac{\partial}{\partial x_1} \left(\frac{1}{k_1^2} u\right) + \left(1 - \alpha_1^2 \frac{1}{k_1^2}\right) u = 0. \tag{3.5}$$

For any fixed $(x_1^0, x_2^0) \in S_1$, $x_2^0 = \phi_1(x_1^0)$, and $r > 0$, denote

$$Q_r = \{(x_1, x_2), \quad |x_1 - x_1^0| < r, \quad |x_2 - x_2^0| < r\}.$$

One may choose R such that $Q_R \subset \Omega$. Consider the transformation in Q_R:

$$x_1' = x_1 - x_1^0, \quad x_2' = x_2 - \phi_1(x_1), \quad u'(x_1', x_2') = u(x_1, x_2), \quad k_1'(x_1', x_2') = k_1(x_1, x_2). \tag{3.6}$$

Using this transformation, Q_R is mapped into a set containing a neighborhood Q_{R_0} of the origin. Without loss of generality, the preimage of Q_{R_0} is assumed to contain $Q_{R/2}$. For simplicity, we shall omit the primes and set (in the new coordinate system)

$$Q_{R_0}^+ = Q_{R_0} \cap \{x_2 > 0\}, \qquad Q_{R_0}^- = Q_{R_0} \cap \{x_2 < 0\}.$$

Consider a more general model problem in Q_{R_0}:

$$\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \frac{\partial}{\partial x_j}(b_j u) + c_j\frac{\partial u}{\partial x_j} + du + f = 0. \tag{3.7}$$

Suppose that $u \in C^\gamma(Q_{R_0}) \cap H^1(Q_{R_0})$. Suppose also that the coefficients $a_{ij}, b_j \in C^\gamma(Q_{R_0}^\pm)$ and $c_j, d, f \in C(Q_{R_0}^\pm)$ have a jump at $x_2 = 0, b_j$, and a_{1j}, a_{2j} are linearly independent, and the principle part of the operator is elliptic in Q_{R_0}, i.e., there is a constant c_0 such that

$$\left|\sum a_{ij}\xi_i\bar{\xi}_j\right| \geq c_0|\xi|^2.$$

Theorem 3.4. *Under the above assumptions, the solution of* (3.7) *satisfies*

$$u \in C^{1,\sigma}(Q_{R_0/4}^\pm) \tag{3.8}$$

for some $0 < \sigma < 1$.

Before proving Theorem 3.4, we make the following remarks.

Theorem 3.4 holds with the same proof in the multidimensional case.

Theorem 3.3 may be proved by using Theorem 3.4. Actually, using the transformation (3.6), equation (3.5) becomes a special case of (3.7) with

$$\begin{aligned}
a_{11} &= \frac{1}{k_1^2}, & a_{12} = a_{21} &= -\frac{1}{k_1^2}\phi_{1x_1}, & a_{22} &= \frac{1}{k_1^2}(\phi_{1x_1}^2 + 1),\\
b_1 &= i\alpha_1\frac{1}{k_1^2}, & b_2 &= -i\alpha_1\phi_{1x_1}\frac{1}{k_1^2}, & f &= 0,\\
c_1 &= i\alpha_1\frac{1}{k_1^2}, & c_2 &= -i\alpha_1\frac{1}{k_1^2}\phi_{1x_1}, & d &= 1 - \frac{\alpha_1^2}{k_1^2}.
\end{aligned}$$

It is easy to check that according to (2.13), the regularity assumptions of Theorem 3.4 are satisfied. By Theorem 3.4, the solution of (3.7) with the above choices of the coefficients satisfies

$$\nabla u \in C^\sigma(Q_{R_0/4}^\pm)$$

for some $0 < \sigma < 1$. Theorem 3.3 then follows.

The proof of (3.8) roughly follows an approach introduced by DiBenedetto, Elliott, and Friedman in [6]. They established the same regularity as in (3.8) for the solution of

$$\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)+B_j\frac{\partial u}{\partial x_j}+Cu=F$$

under the assumptions that

$$a_{ij}\in C^\gamma(Q^\pm_{R_0}),\qquad |\nabla B_j|,\ C,\ F\in L^\infty_{\text{loc}}(Q_{R_0}).$$

However, in our case, since b_j and c_j have a jump across the interface $x_2=0$, the functions $\nabla B_j=\nabla(b_j+c_j)$ and $C=\partial b_j/\partial x_j+d$ may not belong to $L^\infty_{\text{loc}}(Q_{R_0})$. Thus the weak regularity assumptions on the coefficients in our case require more delicate estimates.

Also, in [6], a crucial step for the proof of Hölder continuity of $\partial_{x_1}u$ is to establish the inequality

$$\int_{Q_\rho}|\partial_{x_1}u-(\partial_{x_1}u)_\rho|^2\le CR^{2\beta}\int_{Q_R}|\nabla u|^2+C\left(\frac{\rho}{R}\right)^4\int_{Q_R}|\partial_{x_1}u-(\partial_{x_1}u)_R|^2. \tag{3.9}$$

Again due to the jump of k at $x_2=0$, we are unable to show (3.9) in our case. Fortunately, by some careful estimates, we can obtain a weaker estimate (3.18) below instead of (3.9), which also implies the Hölder continuity of $\partial_{x_1}u$.

We now present our proof of Theorem 3.4. We first assume that $b_2=0$ in equation (3.7). The general case may be reduced to the case of $b_2=0$ by considering the transformation $\tilde u=ue^{\lambda\cdot x}$, where λ_j are determined by

$$\sum_{j=1}^2 a_{ij}\lambda_j+b_i=0.$$

Consider the auxiliary problem

$$\frac{\partial}{\partial x_i}(a^0_{ij}(x_1,x_2)P_{x_j})=0\quad\text{in } Q_{R_0}, \tag{3.10}$$

$$P=u\quad\text{on } \partial Q_{R_0}, \tag{3.11}$$

where $a^0_{ij}=a_{ij}(0,0)$. Because of the jump, equation (3.10) is not an equation with constant coefficients.

Set

$$w_R=\frac{1}{|Q_R|}\int_{Q_R}wdx,\qquad w^\pm_R=\frac{1}{|Q^\pm_R|}\int_{Q^\pm_R}wdx.$$

As shown in [6], the following result holds.

LEMMA 3.5. *For any* $0<\rho<R_0/16$,

$$\int_{Q_\rho}|P_{x_1}-(P_{x_1})_\rho|^2dx\le C\left(\frac{\rho}{R_o}\right)^4\int_{Q_{R_0}}|P_{x_1}-(P_{x_1})_{R_0}|^2dx, \tag{3.12}$$

$$\int_{Q^\pm_\rho}|\nabla P-(\nabla P)^\pm_\rho|^2dx\le C\left(\frac{\rho}{R_0}\right)^4\int_{Q_{R_0}}|P_{x_1}|^2dx. \tag{3.13}$$

Returning to u, we have

$$\begin{aligned}-\frac{\partial}{\partial x_i}(a_{ij}^0(x_1,x_2)(u-P)_{x_j}) = &-\frac{\partial}{\partial x_i}((a_{ij}(x_1,x_2)-a_{ij}^0(x_1,x_2))u_{xj}) \\ &+\frac{\partial}{\partial x_1}(b_1(x_1,x_2)u)+c_j(x_1,x_2)\frac{\partial u}{\partial x_j}+d(x_1,x_2)u+f(x_1,x_2).\end{aligned} \tag{3.14}$$

Lemma 3.6. *There is a constant C that depends on $\|a_{ij}\|_{C^\gamma(Q_{R_0}^\pm)}$, $\|b_1\|_{C^\gamma(Q_{R_0}^\pm)}$, and $\|u\|_{C^\gamma(Q_{R_0})}$ such that*

$$\int_{Q_{R_0}}|\nabla(u-P)|^2 \le CR_0^{2\gamma}\int_{Q_{R_0}}|\nabla u|^2 + CR_0^{2+2\gamma}. \tag{3.15}$$

Proof. Note that $a_{ij}\in C^\gamma(Q_{R_0}^\pm)$ implies that $a_{ij}-a_{ij}^0\in C^\gamma(Q_{R_0})$.

Multiplying (3.14) by $u-P$ and integrating over Q_{R_0}, by using the ellipticity of the operator, we obtain

$$\begin{aligned}\int_{Q_{R_0}}|\nabla(u-P)|^2 \le CR_0{}^{2\gamma}\int_{Q_{R_0}}|\nabla u|^2 &+ \left|\int_{Q_{R_0}}\frac{\partial}{\partial x_j}(b_1u)(u-P)\right| \\ &+\left|\int_{Q_{R_0}}c_j\frac{\partial u}{\partial x_j}(u-P)\right|+\left|\int_{Q_{R_0}}(du+f)(u-P)\right|.\end{aligned} \tag{3.16}$$

Note that no boundary term occurs from the integration by parts since $u-P=0$ on the boundary. We next estimate the last three terms in (3.16). First,

$$\begin{aligned}\left|\int_{Q_{R_0}}c_j(x_1,x_2)\frac{\partial u}{\partial x_j}(u-P)\right| &\le \int_{Q_{R_0}}|c_j(x_1,x_2)|\left|\frac{\partial u}{\partial x_j}\right||u-P| \\ &\le CR_0^{2\gamma}\int_{Q_{R_0}}|\nabla u|^2+R_0^{-2\gamma}\int_{Q_{R_0}}|u-P|^2.\end{aligned}$$

From Poincaré's inequality,

$$\int_{Q_{R_0}}|u-P|^2 \le \frac{1}{2}R_0^2\int_{Q_{R_0}}|\nabla(u-P)|^2,$$

for $R_0^{2-2\gamma}\le 1/2$,

$$\left|\int_{Q_{R_0}}c_j(x_1,x_2)\frac{\partial u}{\partial x_j}(u-P)\right| \le CR_0^{2\gamma}\int_{Q_{R_0}}|\nabla u|^2+\frac{1}{4}\int_{Q_{R_0}}|\nabla(u-P)|^2.$$

Next,

$$\begin{aligned}\left|\int_{Q_{R_0}}(d(x_1,x_2)u+f(x_1,x_2))(u-P)\right| &\le \int_{Q_{R_0}}(|d|\,|u|+|f|)|u-P| \\ &\le CR_0^{2+2\gamma}+\frac{1}{8}\int_{Q_{R_0}}|\nabla(u-P)|^2.\end{aligned}$$

Similarly, integration by parts yields

$$\begin{aligned}\int_{Q_{R_0}} \frac{\partial}{\partial x_1}(b_1 u)(u-P) &= -\int_{Q_{R_0}} (b_1 u - (b_1 u)(0,0)) \frac{\partial (u-P)}{\partial x_1} \\ &\le C R_0^{\gamma} \int_{Q_{R_0}} |\nabla (u-P)| \\ &\le C R_0^{2+2\gamma} + \frac{1}{8}\int_{Q_{R_0}} |\nabla(u-P)|^2.\end{aligned}$$

Estimate (3.15) follows by combining the above estimates. □

The next result is an analogue of Lemma 3.5 on u.

LEMMA 3.7. *For any* $0 < \rho < \eta R$, $0 < \eta < 2^{-4}$, $R \le R_0$,

(3.17)
$$\int_{Q_\rho^{\pm}} |\nabla u - (\nabla u)_\rho^{\pm}|^2 \le C \left[R^{2\gamma} + \left(\frac{\rho}{R}\right)^4 \right] \int_{Q_R} |\nabla u|^2 + C R^{2+2\gamma},$$

(3.18)
$$\begin{aligned}\int_{Q_\rho} |\partial_{x_1} u - (\partial_{x_1} u)_\rho|^2 \le C R^{2\gamma} \int_{Q_R} |\nabla u|^2 + C\left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R|^2 \\ + C R^{2+2\gamma}.\end{aligned}$$

Proof. For any $R > 0$, we decompose u on Q_{R_0} into the sum of P and $P - u$ such that P solves the auxiliary problem (3.10)–(3.11) on Q_R. By using (3.13) and (3.15), we obtain

$$\begin{aligned}\int_{Q_\rho^{\pm}} |\nabla u - (\nabla u)_\rho^{\pm}|^2 &\le \int_{Q_\rho^{\pm}} |\nabla u - \nabla P|^2 + |(\nabla u)_\rho^{\pm} - (\nabla P)_\rho^{\pm}|^2 \\ &\quad + |\nabla P - (\nabla P)_\rho^{\pm}|^2 \\ &\le 2\int_{Q_\rho^{\pm}} |\nabla u - \nabla P|^2 + \int_{Q_\rho^{\pm}} |\nabla P - (\nabla P)_\rho^{\pm}|^2 \\ &\quad + C R^{2\gamma} \int_{Q_R} |\nabla u|^2 + C R^{2+2\gamma} + C\left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} P|^2.\end{aligned}$$

From a simple energy estimate of the auxiliary problem (3.10)–(3.11) on Q_R,

$$\int_{Q_R} |\nabla P|^2 \le C \int_{Q_R} |\nabla u|^2.$$

Thus

$$\int_{Q_\rho^{\pm}} |\nabla u - (\nabla u)_\rho^{\pm}|^2 \le C R^{2\gamma} \int_{Q_R} |\nabla u|^2 + C\left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\nabla u|^2 + C R^{2+2\gamma}.$$

Next, from (3.12) and (3.15),

$$\begin{aligned}\int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_\rho|^2 &\le 2\int_{Q_\rho} |\nabla u - \nabla P|^2 + \int_{Q_\rho^{\pm}} |\nabla P - (\nabla P)_\rho|^2 \\ &\le C\left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} P - (\partial_{x_1} P)_R|^2 + C R^{2\gamma} \int_{Q_R} |\nabla u|^2 \\ &\quad + C R^{2+2\gamma}. \qquad (3.19)\end{aligned}$$

Also,

$$\int_{Q_R} |\partial_{x_1} P - (\partial_{x_1} P)_R|^2 \le 2 \int_{Q_R} |\nabla(P-u)|^2 + \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R|^2. \tag{3.20}$$

Substituting (3.20) into (3.19) yields estimate (3.18). □

LEMMA 3.8. *For any $\varepsilon > 0$, there exist $R_1 \in (0, R_0)$ and $C > 0$ such that for any $0 < \rho < R < R_1$,*

$$\int_{Q_\rho} |\nabla u|^2 \le C \left(\frac{\rho}{R}\right)^{2-\varepsilon} \int_{Q_R} |\nabla u|^2 + C\rho^{2-\varepsilon}, \tag{3.21}$$

where the constants R_1 and C depend on ε.

Proof. Apply Lemma 3.7 with $\rho_i = \eta^i R$, $R < R_1$, and $R_1^{2\gamma} < \eta^4$, where $\eta < 2^{-4}$ is to be determined.

Integrating the inequality

$$|(\nabla u)^\pm_{\rho_{i+1}} - (\nabla u)^\pm_{\rho_i}|^2 \le 2|\nabla u - (\nabla u)^\pm_{\rho_i}|^2 + 2|\nabla u - (\nabla u)^\pm_{\rho_{i+1}}|^2$$

over $Q^\pm_{\rho_{i+1}}$ and dividing by $|Q^\pm_{\rho_{i+1}}|$, we get from (3.17)

$$\begin{aligned} |(\nabla u)^\pm_{\rho_{i+1}} - (\nabla u)^\pm_{\rho_i}|^2 \le{}& \frac{C}{\rho_{i+1}^2}(\rho_i^{2\gamma} + \eta^4) \int_{Q_{\rho_i}} |\nabla u|^2 + \frac{C\rho_i^{2+2\gamma}}{\rho_{i+1}^2} \\ &+ \frac{C}{\rho_{i+1}^2}(\rho_{i-1}^{2\gamma} + \eta^4) \int_{Q_{\rho_{i-1}}} |\nabla u|^2 + \frac{C\rho_{i-1}^{2+2\gamma}}{\rho_{i+1}^2}. \end{aligned} \tag{3.22}$$

Next, apply (3.17) with $R = \rho_i$, $\rho = \rho_{i+1}$:

$$\begin{aligned} \int_{Q^\pm_{\rho_{i+1}}} |\nabla u|^2 &\le \int_{Q^\pm_{\rho_{i+1}}} \left|(\nabla u)^\pm_{\rho_{i+1}}\right|^2 + C(\rho_i^{2\gamma} + \eta^4) \int_{Q_{\rho_i}} |\nabla u|^2 + C\rho_i^{2+2\gamma} \\ &\le |Q^\pm_{\rho_{i+1}}||(\nabla u)^\pm_{\rho_{i+1}}|^2 + C(\rho_i^{2\gamma} + \eta^4) \int_{Q_{\rho_i}} |\nabla u|^2 + C\rho_i^{2+2\gamma}. \end{aligned} \tag{3.23}$$

Observe that

$$\begin{aligned} \left|Q^\pm_{\rho_{i+1}}\right| \left|(\nabla u)^\pm_{\rho_i}\right|^2 &= \frac{|Q^\pm_{\rho_{i+1}}|}{|Q^\pm_{\rho_i}|^2} \left|\int_{Q^\pm_{\rho_i}} \nabla u\right|^2 \\ &\le \frac{|Q^\pm_{\rho_{i+1}}|}{|Q^\pm_{\rho_i}|} \int_{Q^\pm_{\rho_i}} |\nabla u|^2 = \eta^2 \int_{Q^\pm_{\rho_i}} |\nabla u|^2. \end{aligned} \tag{3.24}$$

From (3.22) and (3.24), we have

$$\begin{aligned} \left|Q^\pm_{\rho_{i+1}}\right| \left|(\nabla u)^\pm_{\rho_{i+1}}\right|^2 \le{}& \eta^2 \int_{Q^\pm_{\rho_i}} |\nabla u|^2 + C(\rho_i^{2\gamma} + \eta^4) \int_{Q_{\rho_i}} |\nabla u|^2 \\ &+ C\rho_i^{2+2\gamma} + C(\rho_{i-1}^{2\gamma} + \eta^4) \int_{Q_{\rho_{i-1}}} |\nabla u|^2 + C\rho_{i-1}^{2+2\gamma}. \end{aligned} \tag{3.25}$$

Substituting (3.25) into (3.23), we have

$$\begin{aligned} \int_{Q^\pm_{\rho_{i+1}}} |\nabla u|^2 \le{}& C(\eta^2 + \rho_i^{2\gamma}) \int_{Q_{\rho_i}} |\nabla u|^2 + C(\rho_{i-1}^{2\gamma} + \eta^4) \int_{Q_{\rho_{i-1}}} |\nabla u|^2 \\ &+ C\rho_i^{2+2\gamma} + C\rho_{i-1}^{2+2\gamma}. \end{aligned}$$

By the choice of ρ_i, η^i, and R_1,

$$\rho_i^{2\gamma} < \rho_{i-1}^{2\gamma} < \cdots < \rho_0^{2\gamma} = R^{2\gamma} < R_1^{2\gamma} < \eta^4,$$

we have

$$\int_{Q^\pm_{\rho_{i+1}}} |\nabla u|^2 \le C\eta^2 \int_{Q_{\rho_i}} |\nabla u|^2 + C\eta^4 \int_{Q_{\rho_{i-1}}} |\nabla u|^2 + C\rho_{i-1}^{2+2\gamma}.$$

Iterating with respect to i, we obtain

$$\begin{aligned}\int_{Q^\pm_{\rho_n}} |\nabla u|^2 &\le C\eta^{2(n-1)} \int_{Q_R} |\nabla u|^2 + C(\rho_{n-2}^{2+2\gamma} + \rho_{n-3}^{2+2\gamma}\eta^2 + \cdots + \rho_0^{2+2\gamma}\eta^{2(n-1)})\\ &\le C\eta^{2(n-1)} \int_{Q_R} |\nabla u|^2 + C\rho_0^{2+2\gamma}\\ &\le C_0\eta^{2(n-1)} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma},\end{aligned}$$

where $C > 0$ is independent of n.

We choose η such that

$$\eta^\varepsilon = \frac{1}{C_0}.$$

Then

$$\begin{aligned}\int_{Q^\pm_{\rho_n}} |\nabla u|^2 &\le \eta^{2(n-1)-\varepsilon} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}\\ &\le \eta^{n(2-\varepsilon)-2+(n-1)\varepsilon} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}\\ &\le \left(\frac{\rho_n}{R}\right)^{(2-\varepsilon)} \eta^{(n-1)\varepsilon-2} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}\\ &\le C\left(\frac{\rho_n}{R}\right)^{(2-\varepsilon)} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}, \qquad (3.26)\end{aligned}$$

where we have used $\rho_n/R = \eta^n$ and $|\eta^{(n-1)\varepsilon-2}| \le |\eta^{-2}| \le C$; here the constant C depends on ε.

Assertion (3.21) follows by the following interpolation of (3.26). In fact, for $\rho_{k+1} < \rho < \rho_k$,

$$\begin{aligned}\int_{Q^\pm_\rho} |\nabla u|^2 \le \int_{Q^\pm_{\rho_k}} |\nabla u|^2 &\le C\left(\frac{\rho_k}{R}\right)^{2-\varepsilon} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}\\ &= C\left(\frac{\rho_{k+1}}{\eta R}\right)^{2-\varepsilon} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}\\ &\le C\left(\frac{\rho}{R}\right)^{2-\varepsilon} \int_{Q_R} |\nabla u|^2 + CR^{2+2\gamma}.\end{aligned}$$

The estimate of (3.21) follows from Campanato's lemma [8, p. 86., Lemma 2.1]. □

Lemma 3.9. *For any $0 < \rho < \tilde{R}$, where $\tilde{R} = R_1(\varepsilon)$ is determined in Lemma* 3.8 *for $\varepsilon = \gamma$,*

$$\int_{Q_\rho} |\partial_{x_1} u - (\partial_{x_1} u)_\rho|^2 \leq C \left(\frac{\rho}{R}\right)^{2+\gamma} \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R|^2 + C\rho^{2+\gamma} . \tag{3.27}$$

Proof. Combining (3.18) and (3.21), with $\rho = R$, $R = \tilde{R}$, $0 < \rho < \eta R$, and $\eta < 2^{-4}$, we have

$$\begin{aligned}\int_{Q_\rho} |\partial_{x_1} u - (\partial_{x_1} u)_\rho|^2 &\leq CR^{2\gamma} \int_{Q_R} |\nabla u|^2 + C \left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R|^2 + CR^{2+2\gamma} \\ &\leq CR^{2\gamma} \left[C \left(\frac{R}{\tilde{R}}\right)^{2-\gamma} \int_{Q_{\tilde{R}}} |\nabla u|^2 + CR^{2-\gamma} \right] \\ &\quad + C \left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R| + CR^{2+2\gamma} \\ &\leq CR^{2+\gamma} + C \left(\frac{\rho}{R}\right)^4 \int_{Q_R} |\partial_{x_1} u - (\partial_{x_1} u)_R|.\end{aligned}$$

Hence (3.27) follows from the Campanato lemma.

Using the same arguments as in [6], we can establish (3.27) for any sets Q_ρ and Q_R with center in $Q_{R_0/2}$ since no use has been made of the symmetry or comparative size of $Q^+_{R_0/2}$ and $Q^-_{R_0/2}$. It follows from a Campanato's result that $\partial_{x_1} u$ is Hölder continuous in $Q_{R_0/2}$. The Hölder continuity of $\partial_{x_2} u$ in $\bar{Q}^{\pm}_{R_0/4}$ may also be established by a similar procedure; see, e.g., [6].

Finally, concerning the regularity of the second harmonic field v_α, i.e., solutions of the Helmholtz equation, we have the following general regularity result.

Theorem 3.10. *Under the assumptions* (2.13) *and* (2.14)*, for all but possibly a discrete set of frequencies ω, the problem* (2.34)*,* (2.37)*,* (2.38) *admits a unique solution $v_\alpha \in W^{2,p}(\Omega)$ for $1 < p < \infty$, and the upper bound of its $W^{2,p}$ norm depends on $||\chi^\alpha_{jl}||_{L^\infty(\Omega)}$, $||u_\alpha||_{C^{1,\sigma}(\overline{\Omega_0})}$, and p.*

The proof may be given by combining the regularity results of Theorem 3.3 and the standard elliptic regularity results [9].

Acknowledgments. We thank Professor Avner Friedman for informing us the paper [6]. We also thank the referee and editor for useful comments.

REFERENCES

[1] G. Bao, *Numerical analysis of diffraction by periodic structures: TM polarization*, Numer. Math., to appear.

[2] G. Bao and D. Dobson, *Second harmonic generation in nonlinear optical films*, J. Math. Phys., 35 (1994), pp. 1622–1633.

[3] G. Bao and D. Dobson, *Diffractive optics in nonlinear media with periodic structures*, European J. Appl. Math., 6 (1995), pp. 573–590.

[4] N. Bloembergen, *Nonlinear Optics*, W. A. Benjamin, New York, 1965.

[5] X. Chen and A. Friedman, *Maxwell's equations in a periodic structure*, Trans. Amer. Math. Soc., 323 (1991), pp. 465–507.

[6] E. DiBenedetto, C. M. Elliott, and A. Friedman, *The free boundary of a flow in a porous body heated from its boundary*, Nonlinear Anal., 10 (1986), pp. 879–900.

[7] A. Friedman, *Mathematics in Industrial Problems, Part* 3, Springer-Verlag, Heidelberg, 1990.

[8] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Problems*, Princeton University Press, Princeton, NJ, 1983.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[10] O. A. Ladyzhenskaya and N. N. Ural'ceva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[11] R. Petit, ed., *Electromagnetic Theory of Gratings,* Topics in Current Physics, Vol. 22, Springer-Verlag, Heidelberg, 1980.

[12] E. Popov and M. Nevière, *Surface-enhanced second-harmonic generation in nonlinear corrugated dielectrics: New theoretical approaches*, J. Opt. Soc. Amer. B, 11 (1994), pp. 1555–1564.

[13] R. Reinisch and M. Nevière, *Electromagnetic theory of diffraction in nonlinear optics and surface-enhanced nonlinear optical effects*, Phys. Rev. B, 28 (1983), pp. 1870–1885.

[14] R. Reinisch, M. Nevière, H. Akhouayri, J. Coutaz, D. Maystre, and E. Pic, *Grating enhanced second harmonic generation through electromagnetic resonances*, Opt. Engrg., 27 (1988), pp. 961–971.

[15] Y. R. Shen, *The Principles of Nonlinear Optics*, John Wiley, New York, 1984.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 338–362, March 1997

A SEMILINEAR DIRAC EQUATION IN $H^s(\mathbf{R}^3)$ FOR $s > 1$[*]

M. ESCOBEDO[†] AND L. VEGA[†]

Abstract. Local and global well-posedness for the Cauchy problem associated with the nonlinear Dirac equation

$$i\frac{\partial\psi}{\partial t} + i\alpha\cdot\nabla\psi - m\beta\psi + G(\psi) = 0 \quad \text{in } \mathbf{R}^4$$

are studied in the Sobolev spaces H^s. For regular enough covariant nonlinearities that homogeneous of degree $p \geq 3$, local well-posedness in H^s is proved for $s > \frac{3}{2} - \frac{1}{p-1}$ when p is an odd integer and for $\frac{3}{2} - \frac{1}{p-1} < s < \frac{p-1}{2}$ when p is not an odd integer. If $p > 3$, global well-posedness for small initial data in $H^{s(p)}$, $s(p) = \frac{3}{2} - \frac{1}{p-1}$, is also proved. Local and global well-posedness of the Cauchy problem for the nonlinear Klein–Gordon and wave equations are also considered.

Key words. Dirac equation, well-posedness, Sobolev spaces

AMS subject classifications. 35K22, 35P05

PII. S0036141095283017

Introduction. In this paper, we consider the nonlinear Dirac equation

$$i\frac{\partial\psi}{\partial t} + i\alpha\cdot\nabla\psi - m\beta\psi + G(\psi) = 0 \quad \text{in } \mathbf{R}^4 \tag{0.1}$$

with the following notation:

(i) The unknown ψ is a function from $\mathbf{R}^4$ to $\mathbf{C}^4$ of the variables $(x,t) \in \mathbf{R}^4$ with $x = (x_1, x_2, x_3) \in \mathbf{R}^3$.

(ii)

$$\alpha\cdot\nabla\psi = \sum_{j=1}^{3} \alpha_j \frac{\partial\psi}{\partial x_j},$$

where $(\alpha_j)_{j=1}^3$ are 4×4 matrices defined by

$$\alpha_k = \begin{pmatrix} 0 & \sigma^k \\ \sigma^k & 0 \end{pmatrix}, \quad k = 1,2,3,$$

with

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

(iii) m is a positive real number and β is the 4×4 matrix

$$\beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

Notice that

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij}I,$$

$$\alpha_i\beta + \beta\alpha_i = 0, \qquad \beta^2 = I.$$

[*] Received by the editors March 15, 1995; accepted for publication December 28, 1995.
http://www.siam.org/journals/sima/28-2/28301.html

[†] Departamento de Matemáticas, Universidad del País Vasco, Apartado 644, E-48080 Bilbao, Spain (mtpesmam@lg.ehu.es, mptvegol@lg.ehu.es).

(iv) $G(\cdot)$ is a nonlinear function from $\mathbf{C}^4$ to $\mathbf{C}^4$.

As a simple example of nonlinearity, one can consider $G(u) = |u|^{p-1}u$, where $|\cdot|$ denotes the norm in $\mathbf{C}^4$, or, more generally, $G = (g_1, g_2, g_3, g_4)$, where each of the g_i's is a polynomial in all the components u_j of u. Nevertheless, the more interesting nonlinearities must present some particular structure in order for the equation to be invariant by the Lorentz transformations. These are the so-called covariant nonlinearities. The simplest examples are

$$G(\psi) = |\langle \beta\psi, \psi\rangle|^{\frac{p-1}{2}} \beta\psi + a|\langle \beta\psi, \alpha_5\psi\rangle|^{\frac{p-1}{2}} \alpha_5\beta\psi \quad \text{for } p \geq 1, \tag{0.2}$$

where $a \in \mathbf{C}$, $\alpha_5 = \alpha_1\alpha_2\alpha_3$, and $\langle\, ,\rangle$ denotes the Hermitian product in $\mathbf{C}^4$. The nonlinear Dirac equation has been widely used to build relativistic models of extended particles and extensively studied in the physics literature, in particular, the cubic case $p = 3$ (cf. Finkelstein et al. [8], Rañada [12], Soler [14], and Wakano [16]). Observe that the quantity $\langle\beta\psi, \psi\rangle$ has no definite sign since

$$\langle\beta\psi, \psi\rangle = \psi_1^2 + \psi_2^2 - \psi_3^2 - \psi_4^2.$$

This makes the covariant nonlinearity $G(\psi)$ defined in (0.2) less regular than the noncovariant nonlinearity with the same homogeneity $J(\psi) = |\psi|^{p-1}\psi$. For instance, if $p = 4$, $G(z)$ has less than $\frac{3}{2}$ derivatives in L^∞. We are interested in the Cauchy problem associated with (0.1),

$$\begin{cases} i\dfrac{\partial\psi}{\partial t} + i\alpha\cdot\nabla\psi - m\beta\psi + G(\psi) = 0 & \text{in } \mathbf{R}^4, \\ \psi(x,0) = \psi_0 & \text{in } \mathbf{R}^3, \end{cases} \tag{0.3}$$

in particular, the local and global well-posedness of (0.3) under the weakest possible regularity assumptions. This problem has previously been considered by several authors for different types of nonlinearities. As far as we know, the first well-known result was given by Reed in [13]. He assumed that the nonlinearity $G = (g_1, g_2, g_3, g_4)$ is such that each of the g_i's is a polynomial in all the components u_i, each of whose terms has order p with $p \geq 4$. Then he proved that for all initial data ψ_0 in $H^s(\mathbf{R}^3, \mathbf{C}^4)$ with $s > 3$ small enough, problem (0.3) has a unique solution in $\mathbf{C}(\mathbf{R}, H^3(\mathbf{R}^3, \mathbf{C}^4)) \cap \mathbf{C}^1(\mathbf{R}, H^2(\mathbf{R}^3, \mathbf{C}^4))$ such that its $L^\infty(\mathbf{R}^3, \mathbf{C}^4)$ norm decays in time like $(1+t)^{\frac{3}{2}}$. This result was improved by Dias and Figuera in [5]. They first showed the same global existence result under similar assumptions on G but for $p = 3$ and $s > 2$. Moreover, they proved a global existence result for small initial data in $H^{2+\varepsilon}$ for any $\varepsilon > 0$ and for the cubic covariant nonlinearity (0.2) with $a = 0$. They also obtained an estimate for the decay rate of the $L^\infty(\mathbf{R}^3, \mathbf{C}^4)$ norm as t goes to ∞. Later, in [10], Najman showed in particular the existence of a local solution in the cubic case for any initial data in $H^2(\mathbf{R}^N, \mathbf{C}^4)$. All of these results were far from optimal from the point of view of the local or global existence of solutions. Indeed, as is the case for the Schrödinger, Klein–Gordon, and heat equations, one would expect the following situation. Given an homogeneous nonlinearity G of degree p, there is an exponent $s(p)$ such that, on one hand, the Cauchy problem (0.3) is globally well posed in $H^{s(p)}(\mathbf{R}^3)$ for initial data small enough in the $H^{s(p)}$ norm. On the other hand, the Cauchy problem is locally well posed for any initial data in $H^s(\mathbf{R}^3)$ for $s \geq s(p)$. The value of that critical exponent $s(p)$ is given by the homogeneity of the Cauchy problem and can generally be obtained by scaling arguments.

For instance, let us consider the problem of the existence of solutions to

$$(\mathcal{P})\begin{cases} i\dfrac{\partial\psi}{\partial t}+i\alpha\cdot\nabla\psi+G(\psi)=0 \quad \text{in } \mathbf{R}^4,\\ \psi(x,0)=\psi_0 \quad \text{in } \mathbf{R}^3\end{cases}$$

with $\psi_0\in H^s(\mathbf{R}^3)$. We scale the unknown function ψ in the form

$$\forall\lambda>0,\,\forall t\in\mathbf{R},\,\forall x\in\mathbf{R}^3,\quad \psi_\lambda(t,x)=\lambda^\gamma\psi(\lambda t,\lambda x)$$

with $\gamma+s-\frac{3}{2}=0$ in such a way that, for all $\lambda>0$,

$$||\lambda^\gamma\psi_0(\lambda\cdot)||_{H^s}=\lambda^{-s}||\psi_0||_{L^2}+||\psi_0||_{\dot H^s},$$

where $\dot H^s$ is the homogeneous H^s space defined at the end of this section. Our initial problem is then equivalent to the local existence for each of the following problems:

$$(\mathcal{P}_\lambda)\begin{cases} i\dfrac{\partial\psi_\lambda}{\partial t}+i\alpha\cdot\nabla\psi_\lambda+\lambda^{(s-\frac{3}{2})(p-1)+1}G(\psi_\lambda)=0 \quad \text{in } \mathbf{R}^4,\\ \psi_l(x,0)=\lambda^\gamma\psi_0(\lambda x) \quad \text{in } \mathbf{R}^3.\end{cases}$$

Observe that, at least formally, the smaller $\lambda^{(s-\frac{3}{2})(p-1)+1}$ is, the closer the equation in $(\mathcal{P}_\lambda)$ is to the linear Dirac equation, and hence the easier it should be to find a solution of the Cauchy problem for ψ_λ in H^s and the larger the existence time of that solution should be. However, since on the other hand, the existence time of ψ_λ is an increasing function of λ, the exponent $(s-\frac{3}{2})(p-1)+1$ must be positive or, equivalently,

$$s>\frac{3}{2}-\frac{1}{p-1}.$$

We prove in Theorems I and II below that the value $\frac{3}{2}-\frac{1}{p-1}$ is actually the critical exponent as far as local and global existence are concerned. If we consider the Dirac equation with mass, $m\neq 0$, the same argument can be made if we also scale the mass from m to λm by using Sobolev spaces defined in a slightly different way with a suitable dependence on m.

If the nonlinearity G is regular enough, the local existence in $H^s(\mathbf{R}^N,\mathbf{C}^4)$ for $s>\frac{3}{2}$ is easy to prove using the Haussdorf–Young inequalities (see Lemma 1.3 below) and the Sobolev embedding $H^s(\mathbf{R}^3,\mathbf{C}^4)\subset L^\infty(\mathbf{R}^3,\mathbf{C}^4)$, which holds when $s>\frac{3}{2}$. (Of course, this will not give any decay-rate estimate on $L^\infty(\mathbf{R}^N,\mathbf{C}^4)$ as $t\to\infty$.) As we stated before, this regularity is too high for even covariant nonlinearities. On the other hand, notice that the critical value for s is $\frac{3}{2}-\frac{1}{p-1}$ and therefore smaller than $\frac{3}{2}$. This problem was partially solved in the papers of Reed [13], Dias and Figueira [5], and Moreau [9] by working in spaces involving precise decay rates in time of the spatial Sobolev norms and using the Hausdorff–Young-type inequalities of the linear Dirac group.

Here we use a slightly different argument and obtain local and global solutions under weaker regularity asumptions than in the previous works. As we will see, these are optimal in the case where $p>3$ and, except in the critical case, for $p=3$. This approach is similar to the one used by Cazenave and Weissler [4] for the Schrödinger equation or by Weissler [17] for parabolic equations. Our local results are equivalent to those obtained by Ponce and Sideris in [11] for the nonlinear wave equation. Moreover, the global results for small initial data are new even in that setting. The key idea is to obtain new sharp estimates in $L^{p-1}(\mathbf{R}, L^\infty(\mathbf{R}^N))$ for $p > 3$ (see Theorems 1.5 and 4.2). In order to state our results, let us define, for every $p \geq 1$,

$$s(p) = \max\left\{1, \frac{3}{2} - \frac{1}{p-1}\right\}. \tag{0.4}$$

Observe that if $s(p) < s$, then $\frac{1}{p-1} > \frac{3-2s}{2}$.

Theorem I. *Suppose that $G(u)$ is given by (0.2) for $p \geq 3$. Consider s such that*

$$\begin{cases} s(p) < s < \dfrac{p-1}{2} & \text{if } p > 3 \text{ and is not an odd integer,} \\ s(p) < s & \text{if } p \geq 3 \text{ is an odd integer} \end{cases} \tag{0.5}$$

and let $\gamma > 0$ be such that $\frac{1}{\gamma} \in (\frac{3-2s}{2}, \frac{1}{p-1})$. Then for every $\psi_0 \in H^s(\mathbf{R}^3, \mathbf{C}^4)$, there is a $T^* = T^*(||\psi_0||_{H^s}) > 0$ and a solution ψ to (0.3) such that

$$u \in \mathbf{C}((-T^*, T^*); H^s(\mathbf{R}^3, \mathbf{C}^4)) \cap \mathbf{C}^1((-T^*, T^*); L^2(\mathbf{R}^3, \mathbf{C}^4))$$
$$\cap L^\gamma_{\text{loc}}((-T^*, T^*); L^\infty(\mathbf{R}^3)).$$

Moreover,

(i) *this solution is unique in $L^\infty((-T, T); H^s(\mathbf{R}^3, \mathbf{C}^4)) \cap L^\gamma((-T, T); L^\infty(\mathbf{R}^3))$ for every $T < T^*$;*

(ii) *if $T^* < \infty$, then $||\psi||_{L^\gamma((-T^*,T^*);L^\infty(\mathbf{R}^3))} + ||\psi||_{L^\infty((-T^*,T^*);H^s(\mathbf{R}^3))} = \infty$;*

(iii) *there are a $T \equiv T(||\psi_0||_{H^s}) < T^*(||\psi_0||_{H^s})$ and a neighborhood V of ψ_0 in $H^s(\mathbf{R}^3, \mathbf{C}^4)$ such that for all $0 \leq s' < s$, the map $\psi_0 \longrightarrow \psi(\cdot)$ is continuous from V to $\mathbf{C}((-T, T); H^{s'}(\mathbf{R}^3, \mathbf{C}^4))$.*

Theorem II. *Suppose that G is given by (0.2), $p > 3$, and $s(p) = \frac{3}{2} - \frac{1}{p-1}$. Then for every $\psi_0 \in H^{s(p)}(\mathbf{R}^3, \mathbf{C}^4)$, there are a $T^* = T^*(\psi_0) > 0$ and a solution ψ to (0.3) such that*

$$\psi \in \mathbf{C}((-T^*, T^*); H^{s(p)}(\mathbf{R}^3, \mathbf{C}^4)) \cap \mathbf{C}^1((-T^*, T^*); L^2(\mathbf{R}^3, \mathbf{C}^4))$$
$$\cap L^{p-1}_{\text{loc}}((-T^*, T^*); L^\infty(\mathbf{R}^3)).$$

Moreover,

(i) *this solution is unique in $L^\infty((-T, T); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}((-T, T; L^\infty(\mathbf{R}^3))$ for all $T < T^*$;*

(ii) *if $||\psi_0||_{H^{s(p)}(\mathbf{R}^3,\mathbf{C}^4)}$ is suficiently small, then $T^* = \infty$ and*

$$\psi \in L^\infty((-\infty, \infty); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3)).$$

In this case,

$$\lim_{t \to \pm\infty} ||\psi(t) - W(t)\phi_\pm||_{H^{s(p)}(\mathbf{R}^3)} = 0,$$

where $\{W(t)\}_{t>0}$ stands for the linear Dirac group, $\phi_\pm \in H^{s(p)}(\mathbf{R}^3)$, and

$$\begin{cases} \phi_+ = \psi_0 - i\displaystyle\int_0^\infty W(-\tau)G(\psi)(\tau)d\tau, \\ \phi_- = \psi_0 - i\displaystyle\int_{-\infty}^0 W(-\tau)G(\psi)(\tau)d\tau; \end{cases} \tag{0.6}$$

(iii) *if* $T^* < \infty$*, then* $||\psi||_{L^{p-1}((-T^*,T^*);L^\infty(\mathbf{R}^3))} + ||\psi||_{L^\infty((-T^*,T^*);H^{s(p)}(\mathbf{R}^3))} = \infty$;

(iv) *there are a* $T \equiv T(\psi_0) < T^*(\psi_0)$ *and a neighborhood* V *of* ψ_0 *in* $H^{s(p)}(\mathbf{R}^3, \mathbf{C}^4)$ *such that the map* $\psi_0 \longrightarrow \psi$ *is continuous from* V *to* $\mathbf{C}((-T,T); H^{s'}(\mathbf{R}^3, \mathbf{C}^4))$ *for all* s' *in* $(0, s(p))$*. Finally, if* $||\psi_0||_{H^{s(p)}}$ *is small enough, this is true for* $T = \infty$*.*

Remark 1. Theorem I is true whenever the nonlinearity G satisfies $G(0) = 0$ and is in the Sobolev space $W^{s',\infty}(\mathbf{C}^4, \mathbf{C}^4)$ for some $s' > s$, i.e.,

(a) $d^k G \in L^\infty_{\mathrm{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for $k = 1, \ldots, [s']$, where $[s']$ denotes the integer part of s' and $d^k G$ denotes the kth order differential of G, and

(b) if we set $m = [s']$, then for every $R > 0$,

$$\sup_{|\xi|\le 1, |\zeta| \le 1} \frac{|d^m G(\xi) - d^m(\zeta)|}{|\xi - \zeta|^{s'-m}} = C_R < \infty.$$

On the other hand, Theorem II is true for every nonlinearity G satisfying (a) and (b) that is homogeneous of degree p. Moreover, the same arguments prove that problem (0.3) is globally well posed in $H^1(\mathbf{R}^3)$ for small initial datas if the nonlinearity G is homogeneous of degree $p = 2$ and belongs to $W^{s',\infty}(\mathbf{C}^4, \mathbf{C}^4)$ for some $s' > 1$.

Nevertheless, we observe that there exist only two possibilities in order for a function like $|\langle \beta u, u\rangle|^{\frac{p-1}{2}} \beta u$ to satisfy (a) and (b):

- either p is an odd integer, and then the function $|\langle \beta z, z\rangle|^{\frac{p-1}{2}} \beta z$ is in $\mathbf{C}^\infty$, or
- p is not an odd integer, and then it must be such that $\frac{p-1}{2} > s$.

Remark 2. When $p = 3$, i.e., in the cubic case, Theorem I shows the local existence in $H^s(\mathbf{R}^3, \mathbf{C}^4)$ for all $s > 1$. The local or global existence for the cubic case in $H^1(\mathbf{R}^3, \mathbf{C}^4)$ remains open. On the other hand, Theorem II needs $p > 3$, and therefore we cannot prove global existence, even for small initial data, for the cubic case.

Remark 3. The fact (stated in Theorems I(iii) and II(iv)) that the map $\psi_0 \longrightarrow \psi$ is, in general, only continuous with values in $\mathbf{C}((-T,T); H^{s'}(\mathbf{R}^3, \mathbf{C}^4))$ for all s' in $(0, s)$ and not in $\mathbf{C}((-T,T); H^s(\mathbf{R}^3, \mathbf{C}^4))$ is due to the lack of regularity of the general nonlinearities $G(\cdot)$ that we consider. It is nevertheless easy to check that if the nonlinearity G is regular enough, the maps $\psi_0 \longrightarrow \psi$ defined in Theorems I(iii) and II(iv) are Lipschitz from V into $\mathbf{C}((-T,T); H^s(\mathbf{R}^3, \mathbf{C}^4)) \cap L^\gamma((-T,T), L^\infty(\mathbf{R}^3)$ and $\mathbf{C}((-T,T); H^{s(p)}(\mathbf{R}^3, \mathbf{C}^4)) \cap L^{p-1}((-T,T), L^\infty(\mathbf{R}^3)$, respectively (see Remark 9 in section 3 below).

Remark 4. The size condition on the initial data in Theorem II(ii) is necessary. This follows in particular from the existence results of localized solutions to (0.1) due to Cazenave and Vazquez [3]. These authors proved the existence of solutions of the form $\psi(t,x) = \exp(-i\omega t)\phi(x)$ for any $\omega \in (0, m)$, where $\phi \in \mathbf{C}^1(\mathbf{R}^3, \mathbf{C}^4)$ is such that ϕ and $\nabla\phi$ have an exponential decay as $|x| \to \infty$ (see also Balabane et al. [1]). It is clear that these localized solutions do not belong to the space $L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3))$.

Remark 5. Equation (0.1) has a natural energy given by

$$E(\psi) = -\frac{i}{2} \int_{\mathbf{R}^3} \langle \alpha \cdot \nabla\psi, \beta\psi\rangle dx + \frac{m}{2} \int_{\mathbf{R}^3} \langle \psi, \beta\psi\rangle dx - \int_{\mathbf{R}^3} H(u) dx,$$

where $\nabla H(\psi) = G(\psi)$. This energy E is well defined and of class C^2 on the space $H^{\frac{1}{2}}(\mathbf{R}^3, \mathbf{C}^4)$. It was recently proved by Esteban and Sere in [7] that if $H \in \mathbf{C}^2(\mathbf{R}, \mathbf{R})$, $H(0) = 0$, and for some $\theta > 1$, $H'(x) \ge \theta H(x)$, then E has infinitely many critical points in $H^{\frac{1}{2}}(\mathbf{R}^3, \mathbf{C}^4)$. That gives infinitely many solutions of (0.1) which actually belong to $\bigcap_{2\le q<\infty} W^{1,q}(\mathbf{R}^3, \mathbf{C}^4)$. The space $H^{\frac{1}{2}}(\mathbf{R}^3, \mathbf{C}^4)$ is then the natural one for studying the steady states of (0.1). Nevertheless, it has not yet been possible to use any conservation law to obtain global solutions of (0.1).

In particular, the proof of these results requires estimates on the linear operator

$$A = -i\alpha \cdot \nabla + m\beta. \tag{0.7}$$

This is a self-adjoint operator in $L^2(\mathbf{R}^3, \mathbf{C}^4)$ with domain $H^1(\mathbf{R}^3, \mathbf{C}^4)$. Then iA generates a strongly continuous unitary group in $H^s(\mathbf{R}^3, \mathbf{C}^4)$, which we will denote $(W(t))_{t\in\mathbf{R}}$. Some of the estimates on $W(t)$ given in section 1 have already been used in the study of (0.3) (see, for instance, Dias and Figueira [5], Moreau [9], and Reed [13]).

Remark 6. As we shall see in section 4, similar arguments can be used to study the nonlinear Klein–Gordon and wave equations to obtain results on local and global existence. The estimate in $L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3))$ allows us to prove global existence for small initial data for the critical case established by Ponce and Sideris [11].

Throughout this paper, we will use the following notation:

1. The spaces $L^r(\mathbf{R}^3, \mathbf{C}^4)$ and $H^s(\mathbf{R}^3, \mathbf{C}^4)$ will always be denoted $L^r(\mathbf{R}^3)$ and $H^s(\mathbf{R}^3)$, respectively. Their norms will be denoted $||\cdot||_r$ and $||\cdot||_{H^s}$, respectively.

2. In section 2, we use the Besov spaces $B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4)$ and the homogeneous Besov spaces $\dot{B}^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4)$ for $s > 0$, $p \geq 1$, and $q \geq 1$ reals. They will be denoted $B^s_{p,q}(\mathbf{R}^3)$ and $\dot{B}^s_{p,q}(\mathbf{R}^3)$. We refer the interested reader to the book of Triebel [15] for the definitions and main properties of these spaces. In particular, we will use the following quantities, which are proved to be equivalent to the usual norms (see [15, Theorem, p. 110]),

$$\begin{cases} ||\psi||_{B^s_{p,q}} = ||\psi||_p + ||\psi||_{\dot{B}^s_{p,q}}, \\ ||\psi||_{\dot{B}^s_{p,q}} = \left\{ \int_{\mathbf{R}^3} \dfrac{||\triangle^M_y \psi||^q_p}{|y|^{N+sq}} dy \right\}^{\frac{1}{q}}, \end{cases}$$

with the following notation: M is any integer such that $M > s$, $\triangle_y v(x) = v(x + \frac{y}{2}) - v(x - \frac{y}{2})$ for all $x \in \mathbf{R}^3$ and $y \in \mathbf{R}^3$, and $\triangle^{k+1}_y v \equiv \triangle_y(\triangle^k_y v)$ for all $k \geq 1$. (In the case where $p = q = 2$, this is a direct consequence of the Plancherel identity.)

3. We also use the functional space $L^{r_1}(\mathbf{R}; L^{r_2}(\mathbf{R}^3, \mathbf{C}^4))$ of functions ψ from $\mathbf{R}^4$ to $\mathbf{C}^4$ such that

$$||\psi||_{L^{r_1}(\mathbf{R}; L^{r_2}(\mathbf{R}^3))} \equiv \left\{ \int_{\mathbf{R}} \left(\int_{\mathbf{R}^3} |\psi(t,x)|^{r_2} dx \right)^{\frac{r_1}{r_2}} dt \right\}^{\frac{1}{r_1}} < \infty,$$

where $r_1 \geq 1$ and $r_2 \geq 1$ and where the quantity $||\psi||_{L^{r_1}(\mathbf{R}; L^{r_2}(\mathbf{R}^3))}$ defines a norm. This space will be denoted $L^{r_1}(\mathbf{R}; L^{r_2}(\mathbf{R}^3))$.

4. For $\theta > 0$ and ψ from $\mathbf{R}^4$ to $\mathbf{C}^4$, we denote by $D^\theta_x \psi$ the fractional derivative $(\widehat{D^\theta_x \psi})(\xi) = |\xi|^\theta \hat{\psi}(\xi)$. The homogeneous space $\dot{H}^s$ is defined as the set of measurable functions ψ such that $D^s_x \psi \in L^2$ and

$$||\psi||_{\dot{H}^s} \equiv ||D^s_x \psi||_{L^2}.$$

The rest of the paper is organized as follows. Section 1 covers linear estimates, section 2 covers nonlinear estimates, section 3 contains proof of our main results, and section 4 presents an extension to the Klein–Gordon and wave equations.

1. Linear estimates. Throughout this section, we denote by $\{\psi_j\}_{j\in\mathbf{Z}}$ a smooth partition of unity such that

$$\psi_j(|\xi|) = \psi\left(\frac{|\xi|}{2^j}\right) \quad \text{with supp}\psi \subset (1,2),$$

and we set $\phi_j = \psi_j^2$. For every sufficiently regular function f, we define $S_j f$ and $\tilde{S}_j f$, respectively, by

$$\begin{aligned} \widehat{\left(S_j f\right)}(\xi) &= \phi_j(|\xi|)\hat{f}(|\xi|), \\ \widehat{\left(\tilde{S}_j f\right)}(\xi) &= \psi_j(|\xi|)\hat{f}(|\xi|). \end{aligned}$$

Finally, we define the operator $e^{it\sqrt{m^2-\triangle}}$ acting on a sufficiently regular function f by

$$\mathcal{F}\left(e^{it\sqrt{m^2-\triangle}}f\right)(\xi) = e^{it\sqrt{m^2+|\xi|^2}}\hat{f}(\xi).$$

LEMMA 1.1. (a) *For all $m \geq 0$ and $t > 0$, let*

$$I_j(|x|,t,m) = \int_{\mathbf{R^3}} e^{ix\cdot\xi+it\sqrt{m^2+|\xi|^2}}\psi_j(|\xi|)d\xi \quad \forall x \in \mathbf{R}^3, \quad \forall \xi \in \mathbf{R}^3.$$

Then there is a positive constant C independent of j, t, m, and x such that for all j, t, m, and x,

$$|I_j(|x|,t,m)| \leq C\frac{2^j}{t}(2^{2j}+m^2)^{\frac{1}{2}}. \tag{1.1}$$

(b) *For every $2 \leq p < \infty$, there is a positive constant C_p independent of j, t, and m such that*

$$\begin{cases} \left\|S_j e^{it\sqrt{m^2-\triangle}}f\right\|_p \leq C|t|^{-\alpha(p)}\left\|D_x^{\alpha(p)}(m^2-\triangle)^{\frac{\alpha(p)}{2}}f\right\|_{p'}, \\ \alpha(p) = 2\left(\frac{1}{2}-\frac{1}{p}\right). \end{cases} \tag{1.2}$$

Proof. (a) By a change of variables, we trivially have

$$I_j(|x|,t,m) = \frac{1}{m^3}I_{j\lg_2 m}(|x|m, tm, 1),$$

where $\lg_2 m$ is the logatithm in base 2 of m. Hence we can restrict ourselves to the case where $1 \leq m \leq 2$. Using polar coordinates, we have

$$\begin{aligned} I_j(|x|,t,m) &= \int_0^\infty \psi_j(r)e^{it(m^2+r^2)^{\frac{1}{2}}}r^2\int_{\mathbf{S^2}} e^{ir|x|\cos\theta}\sin\theta\, d\theta \\ &= \int_0^\infty \psi_j(r)e^{it(m^2+r^2)^{\frac{1}{2}}}r\frac{\sin r|x|}{|x|}\,dr. \end{aligned}$$

We shall consider several different cases. First, assume that $|x| \leq 2^{-j+5}$. Then by integration by parts, we have

$$\begin{aligned} |I(|x|,t,m)| &= \left|\int_0^\infty \psi_j(r)\frac{\sin r|x|}{|x|}\frac{(m^2+r^2)^{\frac{1}{2}}}{itr}\frac{d}{dr}\left(e^{it(m^2+r^2)^{\frac{1}{2}}}\right)dr\right| \\ &\leq C\frac{2^j}{t}(2^{2j}+m^2)^{\frac{1}{2}}. \end{aligned}$$

Now consider $|x| > 2^{-j+5}$ and $|t\theta_m - |x|| < 3/4t\theta_m$, where $\theta_m = \sup_{2^j<r<2^{j+1}} r(m^2+r^2)^{-1/2}$. Hence $|x| > t\theta_m/4$, and the estimate immediately follows.

Finally, assume that $|x| > 2^{-j+5}$ and $|t\theta_m - |x|| \geq 3/4t\theta_m$. Then integration by parts gives us

$$\begin{aligned}\frac{1}{|x|}&\left|\int_0^\infty \psi_j(r)r\left(\tfrac{tr}{(m^2+r^2)^{1/2}} \pm |x|\right)^{-1} \tfrac{d}{dr}e^{i(t(m^2+r^2)^{\frac{1}{2}} \pm r|x|)}\,dr\right| \\ &\leq c\frac{2^j}{|x|}\sup_r\left(\frac{tr}{(m^2+r^2)^{\frac{1}{2}}} \pm |x|\right)^{-1} \\ &\leq c\frac{2^j}{|x|}(t\theta_m)^{-1} \leq c\frac{2^j}{t}(2^{2j}+m^2)^{-\frac{1}{2}}.\end{aligned}$$

(b) By definition,

$$\left|S_je^{it\sqrt{m^2-\triangle}}f\right| = \left|\int_{\mathbf{R^3}} e^{ix\cdot\xi+it\sqrt{m^2+|\xi|^2}}\phi_j(|\xi|)\hat{f}(\xi)d\xi\right|,$$

and then, using (a),

$$\left|S_je^{it\sqrt{m^2-\triangle}}f\right| \leq C||\psi_j\hat{f}||_\infty\frac{2^j(m^2+2^{2j})^{\frac{1}{2}}}{|t|} \leq C||\tilde{S}_jf||_{L^1}\frac{2^j(m^2+2^{2j})^{\frac{1}{2}}}{|t|},$$

where $\widehat{(\tilde{S}_jf)} = \psi_j\hat{f}$. Since by Plancherel's theorem we have

$$||S_je^{it\sqrt{m^2-\triangle}}f||_2 \leq C||\psi_j\hat{f}||_2,$$

by Riesz's interpolation theorem, we have, for every $2 \leq p < \infty$,

$$\begin{aligned}\left\|S_je^{it\sqrt{m^2-\triangle}}f\right\|_p &\leq C\left[\frac{2^j(m^2+2^{2j})^{\frac{1}{2}}}{|t|}\right]^{\alpha(p)}||\tilde{S}_jf||_{p'} \\ &\leq C|t|^{-\alpha(p)}\left\|D_x^{\alpha(p)}(m^2-\triangle)^{\frac{\alpha(p)}{2}}S_jf\right\|_{p'} \\ &\leq C|t|^{-\alpha(p)}\left\|D_x^{\alpha(p)}(m^2-\triangle)^{\frac{\alpha(p)}{2}}f\right\|_{p'}. \qquad \square\end{aligned}$$

LEMMA 1.2. *Let $\mathcal{K}_m(t)(f,g)$ be the solution of the initial-value problem (IVP)*

$$\text{(1.3)}\qquad \begin{cases} u_{tt} - \triangle u + m^2u = 0, & x\in\mathbf{R}^3,\quad t\in\mathbf{R},\quad m\geq 0,\\ u(x,0) = f(x), & x\in\mathbf{R}^3,\\ u_t(x,0) = g(x), & x\in\mathbf{R}^3.\end{cases}$$

Then for $2 \leq p < \infty$,

$$\text{(1.4)}\qquad \begin{aligned}||\mathcal{K}_m(t)(f,g)||_{L^p} \leq C|t|^{-\alpha(p)}\Big(&\left\|D_x^{\alpha(p)}(m^2-\Delta)^{\frac{\alpha(p)}{2}}f\right\|_{L^{p'}} \\ &+ \left\|D_x^{\alpha(p)}(m^2-\Delta)^{\frac{\alpha(p)}{2}-1}g\right\|_{L^{p'}}\Big),\end{aligned}$$

where $\alpha(p) = 2(\frac{1}{2}-\frac{1}{p})$ and C is a positive constant independent of m and t.

Proof. By the Littlewood–Paley theorem, we have, for $p \geq 2$,

$$\begin{aligned} ||\mathcal{K}_m(t)(f,g)||_{L^p} &\leq \left\| \left(\sum_{j=-\infty}^{\infty} |S_j \mathcal{K}_m(t)(f,g)|^2 \right)^{\frac{1}{2}} \right\|_{L^p} \\ &\leq \left(\sum_{j=-\infty}^{\infty} ||S_j \mathcal{K}_m(t)(f,g)||_{L^p}^2 \right)^{\frac{1}{2}}. \end{aligned} \tag{1.5}$$

Now using Lemma 1.1(b),

$$\begin{aligned} ||S_j \mathcal{K}_m(t)(f,g)||_{L^p} \leq & \, c|t|^{-\alpha(p)} \left(\left\| D_x^{\alpha(p)} (m^2 - \Delta)^{\frac{\alpha(p)}{2}} \tilde{S}_j f \right\|_{L^{p'}} \right. \\ & \left. + \left\| D_x^{\alpha(p)} (m^2 - \Delta)^{\frac{\alpha(p)}{2-1}} \tilde{S}_j g \right\|_{L^{p'}} \right). \end{aligned} \tag{1.6}$$

Then using (1.5), (1.6), and the fact that $p' \leq 2$,

$$\begin{aligned} ||\mathcal{K}_m(t)(f,g)||_{L^p} \leq c|t|^{-\alpha(p)} & \left(\left\| \left(\sum_j \left| \tilde{S}_j \left(D_x^{\alpha(p)} (m^2 - \Delta)^{\frac{\alpha(p)}{2}} f \right) \right|^2 \right)^{\frac{1}{2}} \right\|_{L^{p'}} \right. \\ & \left. + \left\| \left(\sum_j |\tilde{S}_j (D_x^{\alpha(p)} (m^2 - \Delta)^{\alpha(p)/2-1} g)|^2 \right)^{\frac{1}{2}} \right\|_{L^{p'}} \right). \end{aligned}$$

Using the Littlewood–Paley theorem again, we obtain the desired result. □

LEMMA 1.3. *Let $W(t)\psi_0$ be the solution of*

$$\begin{cases} i\dfrac{\partial \psi}{\partial t} + i\alpha \cdot \nabla \psi - m\beta\psi = 0 \quad \textit{in } \mathbf{R}^4, \\ \psi(x,0) = \psi_0 \quad \textit{in } \mathbf{R}^3. \end{cases} \tag{1.7}$$

Then for $1 < p \leq 2$, there is a positive constant C such that

$$||W(t)\psi_0||_{L^p} \leq C|t|^{-\alpha(p)} ||(D_x^{\alpha(p)} (m^2 - \Delta)^{\frac{\alpha(p)}{2}} \psi_0||_{Lp'},$$

with $\alpha(p) = 2(\frac{1}{2} - \frac{1}{p})$.

Proof. It is well known that $W(t)\psi_0$ solves the IVP (1.3) with $f = \psi_0$ and $g = -\alpha \cdot \nabla\psi_0 - im\beta\psi_0$. Therefore, the desired result follows from Lemma 1.2. □

LEMMA 1.4. *Let $a > 1$, $b > 1$, $\alpha \in (0, N)$, and $\beta \in (0, N)$ such that*

$$\left(\frac{\beta}{N} - \frac{1}{b} \right) \left(\frac{1}{a} - \frac{\alpha}{N} \right) > 0.$$

Then there is a positive constant C such that for every $u \in C^\infty$ tending to zero as $|x|$ goes to ∞, we have

$$||u||_{L^\infty(\mathbf{R}^N)} \leq C||D^\alpha u||_{L^a(\mathbf{R}^N)}^\theta ||D^\beta u||_{L^b(\mathbf{R}^N)}^{1-\theta}, \tag{1.8}$$

with

$$\theta = \frac{\dfrac{\beta}{N} - \dfrac{1}{b}}{\left(\dfrac{\beta}{N} - \dfrac{1}{b} \right) + \left(\dfrac{1}{a} - \dfrac{\alpha}{N} \right)}.$$

Proof. We first suppose that $u \in \mathcal{S}(\mathbf{R}^N)$. Assume first that $\frac{1}{a} - \frac{\alpha}{N} < 0$ and set $f = D_x^\alpha u$. Then for any $\epsilon > 0$, we have

$$\begin{aligned} u = I^\alpha f &= \int f(x-y)\varphi_1\left(\frac{|y|}{\epsilon}\right)\frac{dy}{|y|^{N-\alpha}} + \epsilon^{\alpha-N}\int f(x-y)\varphi_2\left(\frac{|y|}{\epsilon}\right)^{\alpha-N} dy \\ &= I + II, \end{aligned}$$

where $\varphi_1 \in C_0^\infty(\mathbf{R})$, $\varphi_1(r) = 1$ if $|r| \le \frac{1}{2}$, $\varphi_1(r) = 0$ if $r > 1$, and $\varphi_1 + \varphi_2 = 1$.

Hence Hölder's inequality gives

$$|I| \le C\epsilon^{\alpha - N\left(1-\frac{1}{a'}\right)}||f||_{L^a} = c\epsilon^{\alpha-\frac{N}{a}}||f||_{L^a}. \tag{1.9}$$

Set $\omega_\alpha(y) = \varphi_2(|y|)|y|^{\alpha-N}$. Hence $\hat{\omega}_\alpha(\xi) = |\xi|^{-\alpha}\psi(\xi)$ with $\psi \in \mathcal{S}(\mathbf{R^N})$. Therefore, if $\omega_\alpha^\epsilon(y) = \omega_\alpha(y/\epsilon)$, then

$$(D^{\alpha-\beta}\omega_\alpha^\epsilon)\hat{}(\xi) = \epsilon^{N+\beta-\alpha}(\epsilon|\xi|)^{-\beta}\psi(\epsilon\xi).$$

Hence

$$|D^{\alpha-\beta}\omega_\alpha^\epsilon| \le C\epsilon^{\beta-\alpha}\left(1 + \frac{|x|}{\epsilon}\right)^{\beta-N}. \tag{1.10}$$

Now

$$\begin{aligned} |II| &= \epsilon^{\alpha-N}\int D^{\beta-\alpha}fD^{\alpha-\beta}\omega_\alpha^\epsilon \\ &\le \epsilon^{\alpha-N}||D^{\beta-\alpha}f||_{L^b}||D^{\alpha-\beta}\omega_\alpha^\epsilon||_{L^{b'}} \le C\epsilon^{\beta-\frac{N}{b}}||D^{\beta-\alpha}f||_{L^b}, \end{aligned} \tag{1.11}$$

where the last inequality follows from (1.10). Choosing ϵ properly, we obtain (1.8) from (1.9) and (1.11).

If $\frac{1}{a} - \frac{\alpha}{N} > 0$, then by hypothesis, $\frac{1}{b} - \frac{\beta}{N} < 0$ and we can argue as above, using $g = D^\beta u$ instead of f.

The general case for $u \in C^\infty$ tending to zero as $|x|$ goes to ∞ follows by density. □

Theorem 1.5 (Strichartz-type estimates).

(i) *Given* $2 \le p < \infty$, $\frac{1}{q} + \frac{1}{p} = \frac{1}{2}$, *there is a positive constant* C *such that for all* $\psi_0 \in H^{\alpha(p)}(\mathbf{R}^3)$,

$$||W(\cdot)\psi_0||_{L^q(\mathbf{R};L^p(\mathbf{R}^3))} < C||D_x^{\frac{\alpha(p)}{2}}(m^2 - \triangle)^{\frac{\alpha(p)}{4}}\psi_0||_{L^2(\mathbf{R}^3)}, \tag{1.12}$$

where

$$\alpha(p) = 2\left(\frac{1}{2} - \frac{1}{p}\right). \tag{1.13}$$

(ii) *Let* s *and* p *satisfy* (0.5). *Then for all* $\gamma > p - 1$ *such that* $\frac{1}{\gamma} \in (\frac{3-2s}{2}, \frac{1}{p-1})$ *and all* $r' > \gamma$ *such that* $\frac{1}{r'} \in (\frac{3-2s}{2}, \frac{1}{\gamma})$, *there is a positive constant* C *such that for every* $T > 0$ *and all* $\psi_0 \in H^s(\mathbf{R}^3)$,

$$||W(.)\psi_0||_{L^\gamma((-T,T);L^\infty(\mathbf{R}^3))} \le C(1+m)^{\frac{1}{r'}}T^{\frac{1}{\gamma}-\frac{1}{r'}}||\psi_0||_{H^s}. \tag{1.14a}$$

(iii) *Let $p > 3$. Then there is a positive constant C such that*
(1.14b)
$$\forall \psi_0 \in H^{s(p)}(\mathbf{R}^3), \quad ||W(\cdot)\psi_0||_{L^{p-1}((-\infty,\infty);L^\infty(\mathbf{R}^3))} \le C(1+m)^{\frac{1}{p-1}}||\psi_0||_{H^{s(p)}}.$$

Proof. Part (i) follows from standard arguments and is already known; see Brenner [2] and Moreau [9]. Therefore, we only sketch its proof. By looking at the adjoint of the operator $\psi_0 \mapsto W(\cdot)\psi_0$, it is equivalent for us to prove

$$\left\| \int_{-\infty}^{\infty} D^{-\alpha(p)/2}(m^2-\Delta)^{-\alpha(p)/4}W(-t)F(\cdot,t)\,dt \right\|_{L^2} \le C||F||_{L^{q'}(\mathbf{R};L^{p'}(\mathbf{R}^3))}. \tag{1.15}$$

However,

$$\left\| \int_{-\infty}^{\infty} D^{-\alpha(p)/2}(m^2-\Delta)^{-\alpha(p)/4}W(-t)F(x,t)\,dt \right\|_{L^2}^2$$
$$= \int_{-\infty}^{\infty}\int_{\mathbf{R^3}} \overline{F(x,t)} \int_{-\infty}^{\infty} D^{-\alpha(p)}(m^2-\Delta)^{-\alpha(p)/2}W(t-\tau)F(\cdot,\tau)\,d\tau)\,dxdt,$$

and therefore it suffices to prove that

$$\begin{aligned} &\left\| \int_{-\infty}^{\infty} D^{-\alpha(p)}(m^2-\Delta)^{-\alpha(p)/2}W(t-\tau)F(\cdot,\tau)\,d\tau \right\|_{L^q(\mathbf{R};L^p(\mathbf{R}^3))} \\ &\quad \le C||F||_{L^{q'}(\mathbf{R};L^{p'}(\mathbf{R}^3))}. \end{aligned} \tag{1.16}$$

Now by Minkowski's integral inequality and Lemma (1.3), the left-hand side of (1.16) is bounded by

$$C\left\| \int_{-\infty}^{\infty} \frac{d\tau}{|t-\tau|^{\alpha(p)}}\,||F(\cdot,\tau)||_{L_x^{p'}(\mathbf{R^3})} \right\|_{L_t^q}.$$

Then part (i) follows from the Hardy–Littlewood–Sobolev theorem with exponents $\frac{1}{q} = \frac{1}{q'} - (1-\alpha(p))$.

We now prove (ii). Choose $q > 1$ such that

$$2 - s < \frac{1}{q} < \frac{1}{2} + \frac{1}{p-1}$$

and $r > 1$ such that

$$\frac{1}{2} = \frac{1}{q} - \frac{1}{r'}.$$

Observe that

$$\frac{1}{r'} = \frac{1}{q} - \frac{1}{2} < \frac{1}{\gamma}$$

and therefore $r' > \gamma$. We now define

$$\varepsilon = s + \frac{1}{q} - 2,$$

from which we have

$$2 - \frac{1}{q} + \varepsilon \equiv \frac{3}{2} - \frac{1}{r'} + \varepsilon = s,$$

and we set $\theta = \frac{3}{q'} + \varepsilon$. We then have

$$\left(\int_{-T}^{T} ||W(t)\psi_0||_\infty^\gamma dt\right)^{\frac{1}{p-1}} \leq T^{\frac{1}{\gamma} - \frac{1}{r'}} \left(\int_{-T}^{T} ||W(t)\psi_0||_\infty^{r'} dt\right)^{\frac{1}{r'}}.$$

Using Sobolev's embedding since $q'\theta > 3$ and using estimate (i), we get

$$\begin{aligned}\left(\int_{-T}^{T} ||W(t)\psi_0||_\infty^{r'} dt\right)^{\frac{1}{r'}} &\leq C \left(\int_{-T}^{T} ||D_x^\theta W(t)\psi_0||_{q'}^{r'} dt\right)^{\frac{1}{r'}} \\ &\leq C(1+m)^{\frac{\alpha(q')}{2}} ||\psi_0||_{H^{\theta+\alpha(q')}}.\end{aligned}$$

By the choice of θ and the definition $\alpha(q') \equiv \frac{2}{r'}$, we have $\theta + \alpha(q') \equiv s$, and (ii) follows.

To prove (iii), suppose that $p > 3$ and $s = s(p)$. In this case we estimate the L^∞-norm using the Gagliardo–Nirenberg-type inequality given by Lemma 1.4 with exponents

$$\begin{aligned} &a = 2, \qquad \alpha = s(p) = \frac{3}{2} - \frac{1}{p-1}, \qquad \frac{1}{b} = \frac{1}{4} - \frac{1}{2(p-1)}, \\ &\beta = 1 - \frac{2}{p-1}, \qquad \theta = \frac{p-3}{p+1}, \qquad 1 - \theta = \frac{4}{p+1}, \end{aligned}$$

we obtain

$$\begin{aligned} &||W(t)\psi_0||_{L^{p-1}((\mathbf{R});L^\infty(\mathbf{R}^3))} \\ &\quad \leq C\left(\int_{-\infty}^{\infty} ||D_x^{s(p)}W(t)\psi_0||_{L^2}^{\theta(p-1)} ||D_x^\beta W(t)\psi_0||_{L^b}^{(1-\theta)(p-1)} dt\right)^{\frac{1}{p-1}} \\ &\quad \leq C||D_x^{s(p)}W(t)\psi_0||_{L^\infty((\mathbf{R});L^2)}^\theta \left(\int_{-\infty}^{\infty} ||D_x^\beta W(t)\psi_0||_{L^b}^{(1-\theta)(p-1)} dt\right)^{\frac{1}{p-1}}. \end{aligned}$$

We now observe that the pairs $(\infty, 2)$ and $((1-\theta)(p-1), b)$ are admissible exponents; in particular,

$$\frac{1}{(1-\theta)(p-1)} + \frac{1}{b} = \frac{p+1}{4(p-1)} + \frac{1}{4} - \frac{1}{2(p-1)} = \frac{1}{2}.$$

Therefore, by (i),

$$||D_x^{s(p)}W(t)\psi_0||_{L^\infty((\mathbf{R});L^2(\mathbf{R}^3))} \leq C||D_x^{s(p)}\psi_0||_{L^2(\mathbf{R}^3)} \leq C||\psi_0||_{H^{s(p)}},$$

where we have used $\alpha(2) = 0$, and

$$\begin{aligned}\left(\int_{-\infty}^{\infty} ||D_x^\beta W(t)\psi_0||_{L^b(\mathbf{R}^3)}^{(1-\theta)(p-1)} dt\right)^{\frac{1}{p-1}} &\leq C(1+m)^{\frac{(1-\theta)\alpha(b)}{2}} ||D_x^\beta \psi_0||_{H^{\alpha(b)}}^{1-\theta} \\ &\leq C(1+m)^{\frac{1}{p-1}} ||\psi_0||_{H^{s(p)}}^{1-\theta}\end{aligned}$$

since $\beta + \alpha(b) = s(p)$. The result follows since $\alpha(b) = \frac{1}{2} + \frac{1}{p-1}$. □

Remark 7. The restriction $p > 3$ in Theorems 1.5(ii) and II comes then from the use of Sobolev's embedding when $s > s(p)$ and the Gagliardo–Nirenberg inequality when $s = s(p)$.

2. Nonlinear estimates. This section is devoted to the proof of a kind of fractional chain rule on the Besov spaces $B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4)$. Let us recall here that on these spaces we use the norm

$$\begin{cases} ||\psi||_{B^s_{p,q}} = ||\psi||_p + ||\psi||_{\dot{B}^s_{p,q}}, \\ ||\psi||_{\dot{B}^s_{p,q}} = \left\{ \int_{\mathbf{R}^4} \frac{||\triangle^M_y \psi||^q_p}{|y|^{N+sq}} dy \right\}^{\frac{1}{q}} \end{cases}$$

for any integer $M > s$ and where $\triangle_y v(\cdot) = v(\cdot + \frac{y}{2}) - v(\cdot - \frac{y}{2})$. We are mainly interested in the application of Proposition 2.1 and Corollary 2.2 below to the case where $p = q = 2$. In this case, the Plancherel identity shows that $||\cdot||_{B^s_{p,q}}$ and $||\cdot||_{\dot{B}^s_{p,q}}$ are actually equivalent to the usual norms on $B^s_{p,q}$ and $\dot{B}^s_{p,q}$, respectively.

Proposition 2.1. *Let be $s > 1$ and $m = [s]$. Suppose that a given function ϕ belongs to $W^{s',\infty}_{\text{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for some $s' \in (s, [s]+1)$ and satisfies $\phi(0) = 0$. Then for any $\psi \in B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4) \cap L^\infty(\mathbf{R}^3, \mathbf{C}^4)$, $\phi(\psi) \in B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4)$ and, moreover,*

$$\begin{aligned} ||\phi(\psi)||_{B^s_{p,q}} \leq \sup_{|z| \leq ||\psi||_\infty} |d\phi(z)|||\psi||_p + \sum_{k=1}^{[s]} \sup_{|z| \leq ||\psi||_\infty} |d^k \phi(z)|||\psi||^{k-1}_\infty ||\psi||_{\dot{B}^s_{p,q}} \\ + C(||\psi||_\infty) ||\psi||_{\dot{B}^s_{p,q}} ||\psi||^{s'-1}_\infty, \end{aligned}$$

where for every $R > 0$, the constant C_R is defined by

$$C(R) = \sup_{|\xi| \leq R, |\zeta| \leq R} \frac{|d^m G(\xi) - d^m G(\zeta)|}{|\xi - \zeta|^{s'-m}} \tag{2.1}$$

and $d^k \phi$ is the differential of order k of ϕ.

Proof. This proposition is esentially proved in Escobedo [6]. For the sake of completeness, we will give the details of the proof for the case where $s \in (1, 2)$. The general case can easily be completed by using the results in [6].

First, since $\phi(0) = 0$,

$$||\phi(\psi)||_p \leq \sup_{|z| \leq ||\psi||_\infty} |d\phi(z)|||\psi||_p.$$

We now estimate $||\phi(\psi)||_{\dot{B}^s_{p,q}}$. For this observe that

$$\triangle^2_y \phi(\psi) = \triangle_y(\triangle_y \phi(\psi))$$

and then, for every $x \in \mathbf{R}^3$,

$$\begin{aligned} \triangle_y \phi(\psi)(x) &= \phi\left(\psi\left(x + \tfrac{y}{2}\right)\right) - \phi\left(\psi\left(x - \tfrac{y}{2}\right)\right) \\ &= \int_0^1 d\phi\left(\theta\psi\left(x + \frac{y}{2}\right) + (1 - \theta\psi\left(x - \frac{y}{2}\right)\right) \triangle_y \psi(x) d\theta, \end{aligned}$$

from which we have

$$\begin{aligned}\triangle_y^2\phi(\psi)(x) &= \int_0^1 \triangle_y\left[d\phi\left(\theta\psi\left(x+\frac{y}{2}\right)+(1-\theta)\psi\left(x-\frac{y}{2}\right)\right)\triangle_y\psi(x)\right]d\theta\\ &= \int_0^1 d\phi\left(\theta\psi\left(x+\frac{y}{2}\right)+(1-\theta)\psi\left(x-\frac{y}{2}\right)\right)\triangle_y^2\psi(x)d\theta\\ &\quad+\int_0^1 \triangle_y\left(d\phi\left(\theta\psi\left(x+\frac{y}{2}\right)+(1-\theta)\psi\left(x-\frac{y}{2}\right)\right)\right)\triangle_y\psi(x)d\theta\\ &\equiv I_1+I_2.\end{aligned}$$

First, we estimate I_1, which is the easiest. Indeed, since

$$|I_1| \le \sup_{|z|\le||\psi||_\infty} |d\phi(z)||\triangle_y^2\psi|,$$

we deduce

$$\left\{\int_{\mathbf{R}^4}\frac{||I_1||_p^q}{|y|^{N+sq}}dy\right\}^{\frac{1}{q}} \le \sup_{|z|\le||\psi||_\infty}|d\phi(z)|||\psi||_{\dot{B}^s_{p,q}}.$$

Now consider I_2. Since

$$\begin{aligned}&\triangle_y\Big(d\phi\left(\theta\psi\left(x+\frac{y}{2}\right)+(1-\theta)\psi\left(x-\frac{y}{2}\right)\right)\Big)\\ &\qquad= d\phi(\theta\psi(x+y)+(1-\theta)\psi(x))-d\phi(\theta\psi(x)+(1-\theta)\psi(x-y)),\end{aligned}$$

we have, using the hypothesis on ϕ,

$$\begin{aligned}&\Big|\triangle_y\left(d\phi\left(\theta\psi\left(x+\frac{y}{2}\right)+(1-\theta)\psi\left(x-\frac{y}{2}\right)\right)\right)\Big||\triangle_y\psi|\\ &\quad\le C(||u||_\infty)|\theta\psi(x+y)+(1-\theta)\psi(x)-\theta\psi(x)-(1-\theta)\psi(x-y)|^{s'-1}|\triangle_y\psi|\\ &\quad\le C(||u||_\infty)C'_s|\triangle_y\psi|(|\psi(x+y)-\psi(x)|^{s'-1}-|\psi(x)-\psi(x-y)|^{s'-1})\\ &\quad= C(||u||_\infty)C'_s|\triangle_y\psi|(|\tau_y\triangle_y\psi(x)|^{s'-1}+|\tau_{-y}\triangle_y\psi(x)|^{s'-1}),\end{aligned}$$

where for every $x \in \mathbf{R}^3$ and $y \in \mathbf{R}^3$, $\tau_y\psi(x) = \psi(x+\frac{y}{2})$. We now take $\varepsilon > 0$ such that $s' > s + (1 - \frac{s}{s+\varepsilon})$ and define $\rho = s + \varepsilon$. Then

$$s' > s+\left(1-\frac{s}{s+\varepsilon}\right) \Longrightarrow s'-1 > s\left(1-\frac{1}{s+\varepsilon}\right) \equiv s\left(1-\frac{1}{\rho}\right) \Longrightarrow \rho's' > s.$$

Using Hölder's inequality

$$\begin{aligned}||I_2||_p &\le C(||u||_\infty)C'_s||\triangle_y\psi||_{\rho p}|||\tau_y\triangle_y\psi|^{s'-1}+|\tau_{-y}\triangle_y\psi|^{s'-1}||_{\rho' p}\\ &\le 2C(||u||_\infty)C'||\triangle_y\psi||_{\rho p}||\triangle_y\psi||^{s'-1}_{(s'-1)\rho' p},\end{aligned}$$

where C' is a constant depending only on s', p, and ρ. Again using Hölder's inequality,

$$\begin{aligned}&\left\{\int_{\mathbf{R}^4}\frac{||I_2||_p^q}{|y|^{N+sq}}dy\right\}^{\frac{1}{q}}\\ &\quad\le 2C(||u||_\infty)C'\left\{\int_{\mathbf{R}^4}\frac{||\triangle_y\psi||_{\rho p}^{\rho q}}{|y|^{N+sq}}dy\right\}^{\frac{1}{\rho q}}\left\{\int_{\mathbf{R}^4}\frac{||\triangle_y\psi||^{(s'-1)\rho' q}_{(s'-1)\rho' p}}{|y|^{N+sq}}dy\right\}^{\frac{1}{\rho' q}}.\end{aligned}$$

Since $s < r$ and $s < \rho'(s'-1)$, by the definition of the homogeneous norms that we are using,

$$\left\{ \int_{\mathbf{R}^4} \frac{||I_2||_p^q}{|y|^{N+sq}} dy \right\}^{\frac{1}{q}} \leq ||\psi||_{\dot{B}^{\frac{s}{\rho}}_{\rho p, \rho q}} ||\psi||^{(s'-1)}_{\dot{B}^{\frac{s}{(s'-1)\rho'}}_{(s'-1)\rho' p, (s'-1)\rho' q}} .$$

Finally, observe that since $\rho > 1$ and $\rho'(s'-1) > 1$, we have for any $M > s$ that

$$||\triangle_y \psi||_{\rho p}^{\rho q} = \left\{ \int_{\mathbf{R}^3} |\triangle_y^M \psi(x)|^{\rho p} dx \right\}^{\frac{q}{p}} \leq ((M+1)||\psi||_\infty^{\rho - 1})^q \left\{ \int_{\mathbf{R}^3} |\triangle_y^M \psi(x)|^p dx \right\}^{\frac{q}{p}} .$$

We easily deduce that for some positive constant C depending only on ρ, p, q, and s, we have

$$\begin{aligned} ||\psi||_{\dot{B}^{\frac{s}{\rho}}_{\rho p, \rho q}} &\leq C ||\psi||_\infty^{1 - \frac{1}{\rho}} ||\psi||^{\frac{1}{\rho}}_{\dot{B}^s_{p,q}} , \\ ||\psi||^{(s'-1)}_{\dot{B}^{\frac{s}{(s'-1)\rho'}}_{(s'-1)\rho' p, (s'-1)\rho' q}} &\leq C ||\psi||_\infty^{(s'-1) - \frac{1}{\rho'}} ||\psi||^{\frac{1}{\rho'}}_{\dot{B}^s_{p,q}} , \end{aligned}$$

and Proposition 1.2 follows. □

Corollary 2.2. *Let be $s > 1$ and $m = [s]$. Suppose that ϕ is a given homogeneous function from $\mathbf{C}^4$ to $\mathbf{C}^4$ of degree p and belongs to $W^{s',\infty}_{\text{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for some $s' > s$. Then there is a positive constant C depending only on s, s', p, and q such that for any $\psi \in B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4) \cap L^\infty(\mathbf{R}^3, \mathbf{C}^4)$, $\phi(\psi) \in B^s_{p,q}(\mathbf{R}^3, \mathbf{C}^4)$ and, moreover,*

$$||\phi(\psi)||_{B^s_{p,q}} \leq C \left[C(1) + \sup_{|z| \leq 1} |d\phi(z)| + \sum_{k=1}^{[s]} \sup_{|z| \leq 1} |d^k \phi(z)| \right] ||\psi||_{B^s_{p,q}} ||\psi||_\infty^{p-1},$$

where $C(1)$ is given by (2.1) *for $R = 1$.*

Proof. If ϕ is homogeneous of degree p, its differential of order $k < [s']$ is homegeneous of degree $p - k$. On the other hand, if $m = [s]$,

$$\begin{aligned} C(R) &= \sup_{|\xi| \leq R, |\zeta| \leq R} \frac{|d^m \phi(\xi) - d^m \phi(\zeta)|}{|\xi - \zeta|^{s' - m}} \\ &\equiv \sup_{|\xi| \leq R, |\zeta| \leq R} \frac{R^{p-1} |d^m \phi(\frac{\xi}{R}) - d^m \phi(\frac{\zeta}{R})|}{R^{s'-1} |\frac{\xi}{R} - \frac{\zeta}{R}|^{s'-m}} \\ &= R^{p - s'} \sup_{|\xi| \leq 1, |\zeta| \leq 1} \frac{|d^m \phi(\xi) - d^m \phi(\zeta)|}{|\xi - \zeta|^{s'-m}} . \end{aligned}$$

The result then follows by Proposition 2.1. □

Remark 8. If we consider the function G defined in (0.2) with $p \geq 3$ and s satisfying (0.5) or $s = s(p)$ when $p > 3$, then G is homogeneous of degree p. If $p \geq 3$ is an odd integer, G is C^∞ and then belongs to $W^{s',\infty}_{\text{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for all s'. If p is not an odd integer, then G belongs to $W^{s',\infty}_{\text{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for $s' = \frac{p-1}{2}$, which by (0.5) is greater than s. Corollary 2.2 can then be applied to $G(\psi)$, and this will be used in Lemma 3.1 below.

3. Proof of the main results. As stated in the introduction, the operator iA generates a strongly continuous unitary group in H^s, which we denote by $W(t)$. We can write then the Cauchy problem (0.3) in the following integral form:

$$\psi(t) = W(t)\psi_0 - i\int_0^t W(t-\tau)G(\psi)d\tau. \tag{3.1}$$

We assume throughout this section that G is given by (0.2) with $p \geq 3$ and that s satisfies (0.5) or $s = s(p)$ and $p > 3$.

LEMMA 3.1.

(i) *Suppose that $p > 3$. Then there is $C > 0$ such that for every $T \in (0, \infty]$ and all $\psi \in L^\infty((-T,T); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}((-T,T); L^\infty(\mathbf{R}^3))$, we have*

$$\int_0^t W(t-\tau)G(\psi)d\tau \in L^{p-1}((-T,T); L^\infty(\mathbf{R}^3))$$

and

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)G(\psi)d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\ &\qquad \leq C||\psi||_{L^\infty((-T,T);H^{s(p)}(\mathbf{R}^3))}||\psi||^{p-1}_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}.\end{aligned}$$

(ii) *Suppose that p and s satisfy (0.5) and assume that $\gamma > (p-1)$ is such that $s > \frac{3}{2} - \frac{1}{\gamma}$. Suppose that $\psi \in L^\infty((-T,T); H^{s(p)}(\mathbf{R}^3)) \cap L^\gamma((-T,T); L^\infty(\mathbf{R}^3))$. Then*

$$\int_0^t W(t-\tau)G(\psi)d\tau \in L^{p-1}((-T,T); L^\infty(\mathbf{R}^3))$$

and, moreover,

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)G(\psi)d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\ &\qquad \leq CT^{\frac{1}{p-1}-\frac{1}{\gamma}}||\psi||_{L^\infty((-T,T);H^{s(p)}(\mathbf{R}^3))}||\psi||^{p-1}_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\end{aligned}$$

for every r' such that $\frac{1}{r'} \in (2-s, \frac{1}{\gamma})$.

Proof. First, we prove (i). Using Theorem 1.5, we have

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)G(\psi)(\tau)d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\ &\quad \leq \left\|\int_{-T}^T |W(t-\tau)G(\psi)(\tau)|d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\ &\quad \leq \int_{-T}^T ||W(t-\tau)G(\psi)(\tau)||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}d\tau \leq C\int_{-T}^T ||G(\psi)(\tau)||_{H^{s(p)}}d\tau,\end{aligned}$$

where in the last inequality we used Theorem 1.5(iii). Now by Corollary 2.2,

$$\int_{-T}^T ||G(\psi)(\tau)||_{H^{s(p)}}d\tau \leq C||\psi||_{L^\infty((-T,T);H^{s(p)}(\mathbf{R}^3))}\int_{-T}^T ||\psi(\tau)||^{p-1}_\infty d\tau.$$

The proof of (ii) is similar. Indeed,

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)G(\psi)d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\&\quad\le\left\|\textstyle\int_{-T}^T |W(t-\tau)G(\psi)|d\tau\right\|_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\\&\quad\le \textstyle\int_{-T}^T ||W(t-\tau)G(\psi)||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}d\tau\le CT^{\frac{1}{p-1}-\frac{1}{\gamma}}\int_{-T}^T ||G(\psi)||_{H^s}d\tau,\end{aligned}$$

where in the last inequality we used Theorem 1.5(ii). Again using Corollary 2.2,

$$\int_{-T}^T ||G(\psi)||_{H^s}d\tau\le C||\psi||_{L^\infty((-T,T);H^s(\mathbf{R}^3))}\int_{-T}^T ||\psi(\tau)||_\infty^{p-1}d\tau.$$

Finally, (ii) follows using Hölder's inequality since $\gamma>p-1$. □

LEMMA 3.2.

(i) *Let* $\gamma>p-1$ *and* $T\in(0,\infty]$. *If* ψ_1 *and* ψ_2 *are in* $L^\infty((0,T);L^2(\mathbf{R}^3))\cap L^\gamma((0,T);L^\infty(\mathbf{R}^3))$ *for some* $T\in\mathbf{R}$, *then*

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)(G(\psi_1)-G(\psi_2))d\tau\right\|_{L^\infty((0,T);L^2(\mathbf{R}^3))}\\&\quad\le CT^{1-\frac{(p-1)}{\gamma}}||\psi_1-\psi_2||_{L^\infty((0,T);L^2(\mathbf{R}^3))}\left(||\psi_1||^{p-1}_{L^\gamma((0,T);L^\infty(\mathbf{R}^3))}+||\psi_2||^{p-1}_{L^\gamma((0,T);L^\infty(\mathbf{R}^3))}\right).\end{aligned}$$

(ii) *For every* $T\in(0,\infty]$, *if* ψ_1 *and* ψ_2 *are in* $L^\infty((-T,T);L^2(\mathbf{R}^3))\cap L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))$, *then*

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)(G(\psi_1)-G(\psi_2))d\tau\right\|_{L^\infty((-T,T);L^2(\mathbf{R}^3))}\\&\quad\le C||\psi_1-\psi_2||_{L^\infty((-T,T);L^2(\mathbf{R}^3))}\left(||\psi_1||^{p-1}_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}+||\psi_2||^{p-1}_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}\right).\end{aligned}$$

Proof. We prove (i); the proof of (ii) is similar. By our linear estimates, for all $t\in(0,T)$,

$$\begin{aligned}&\left\|\int_0^t W(t-\tau)(G(\psi_1(\tau))-G(\psi_2(\tau)))d\tau\right\|_{L^2(\mathbf{R}^3)}\\&\qquad\le\int_0^t ||(G(\psi_1(\tau))-G(\psi_2(\tau)))||_{L^2(\mathbf{R}^3)}d\tau.\end{aligned}$$

Moreover, since $\frac{p-1}{2}>s>1$,

$$|G(\psi_1)-G(\psi_2)|\le C_p|\psi_1-\psi_2|(|\psi_1|^{p-1}+|\psi_2|^{p-1}),$$

from which we have, for all $t\in(0,T)$,

$$\begin{aligned}&\int_0^t ||(G(\psi_1(\tau))-G(\psi_2(\tau)))||_{L^2(\mathbf{R}^3)}d\tau\\&\quad\le C_p||\psi_1-\psi_2||_{L^\infty((0,T);L^2(\mathbf{R}^3))}\int_0^t\left(||\psi_1(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)}+||\psi_2(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)}\right)d\tau,\end{aligned}$$

and the result follows. □

LEMMA 3.3. *Let be $\psi_0 \in H^s(\mathbf{R}^3)$ and $\gamma \geq p-1$. Suppose that $\psi \in L^\gamma((-T,T);L^\infty(\mathbf{R}^3)) \cap L^\infty((-T,T);H^s(\mathbf{R}^3))$ is a solution of equation* (3.1). *Then $\psi \in \mathbf{C}([-T,T];H^s(\mathbf{R}^3))$. Furthermore, if ψ_1 and ψ_2 are two such solutions, then $\psi_1=\psi_2$.*

Proof. The first statement follows from the facts that $W(t)$ is a strongly continuous group on $H^s(\mathbf{R}^3)$ and that $G(\psi) \in L^1((-T,T);H^s(\mathbf{R}^3))$. The proof of the second statement is by now classical. We write it down for the sake of completeness following Cazenave and Weissler [4]. Suppose that $\psi_1(t) \neq \psi_2(t)$ for some $t \in (-T,T)$. Let $t_0 = \inf\{t \in (-T,T); \psi_1(t) \neq \psi_2(t)\}$. Since ψ_1 and ψ_2 are continuous, t_0 is well defined and $\psi_1(t_0)=\psi_2(t_0)$. Moreover, the functions $\phi_1(t)=\psi_1(t+t_0)$ and $\phi_2(t)=\psi_2(t+t_0)$ are two solutions of equation (3.1) on the interval $(0,T-t_0)$. Then by Lemma 3.2, for all $t \in (t_0,T)$,

$$\begin{aligned} ||\psi_1-\psi_2||_{L^\infty((t_0,t);L^2(\mathbf{R}^3))} &= \left\| \int_0^t W(t-\tau)(G(\psi_1)-G(\psi_2))d\tau \right\|_{L^\infty((t_0,t);L^2(\mathbf{R}^3))} \\ &\leq C(t-t_0)^{1-\frac{(p-1)}{\gamma}} ||\psi_1-\psi_2||_{L^\infty((t_0,t);L^2(\mathbf{R}^3))} \\ &\quad \cdot \left(||\psi_1||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))} + ||\psi_2||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))} \right). \end{aligned}$$

For $t>t_0$ but sufficiently close to t_0, it follows that

$$C(t-t_0)^{1-\frac{(p-1)}{\gamma}} \left(||\psi_1||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))} + ||\psi_2||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))} \right) < 1.$$

(Observe that even if $\gamma = p-1$, since ψ_1 and ψ_2 belong to $L^\gamma((-T,T);L^\infty(\mathbf{R}^3))$, we have that $||\psi_1||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))}$ and $||\psi_2||^{p-1}_{L^\gamma((t_0,t);L^\infty(\mathbf{R}^3))}$ tend to 0 as $t \to t_0$.) This implies that $||\psi_1-\psi_2||_{L^\infty((t_0,t);L^2(\mathbf{R}^3))}=0$, which contradicts the choice of t_0 and thus proves $\psi_1(t)=\psi_2(t)$ for all $t \in (-T,T)$. □

Proof of Theorem I. Let $\gamma > p-1$ such that $\frac{3}{2}-\frac{1}{\gamma}<s$. For $T>0$ and $M>0$, we define

$$\mathbf{X}(T,M) = \{\psi \in L^\infty((-T,T);H^s(\mathbf{R}^3)) \cap L^\gamma((-T,T);L^\infty(\mathbf{R}^3)); |||\psi||| \leq M\},$$

where

$$|||\psi||| = \sup_{t\in(-T,T)} ||\psi||_{H^s(\mathbf{R}^3)} + ||\psi||_{L^\gamma((-T,T);L^\infty(\mathbf{R}^3))}.$$

The set $\mathbf{X}(T,M)$ endowed with the metric

$$d(\psi_1,\psi_2) = ||\psi_1-\psi_2||_{L^\infty((-T,T);L^2(\mathbf{R}^3))}$$

is a complete metric space. We want to find conditions on T and M which imply that the map $\mathcal{F}$, given by

$$\mathcal{F}(\psi) = W(t)\psi_0 - i\int_0^t W(t-\tau)G(\psi)d\tau,$$

is a strict contraction on $\mathbf{X}(T,M)$. By Lemma 3.1, there are positive constants C_1 and C_2 depending only on s, p, and γ such that for all $\psi \in \mathbf{X}(T,M)$,

$$|||\mathcal{F}(\psi)||| \leq C_1||\psi_0||_{H^s(\mathbf{R}^3)} + C_2T^{1+\frac{2-p}{\gamma}-\frac{1}{r'}}M^p.$$

Then if

$$C_1||\psi_0||_{H^s(\mathbf{R}^3)} + C_2 T^{1+\frac{2-p}{\gamma}-\frac{1}{r'}} M^p \leq M, \tag{3.2}$$

$\mathcal{F}(\psi) \in \mathbf{X}(T, M)$. On the other hand, by Lemma 3.2, there is a positive constant C_3 depending only on s, p, and γ such that for all $\psi_1 \in \mathbf{X}(T, M)$ and $\psi_2 \in \mathbf{X}(T, M)$,

$$d(\mathcal{F}(\psi_1), \mathcal{F}(\psi_2)) < C_3 T^{1-\frac{(p-1)}{\gamma}} M^{p-1}.$$

Then if

$$C_3 T^{1-\frac{(p-1)}{\gamma}} M^{p-1} < 1, \tag{3.3}$$

$\mathcal{F}$ is a strict contraction. Therefore, if T and M satisfy (3.2) and (3.3), $\mathcal{F}$ is a strict contraction on $\mathbf{X}(T, M)$. Thus given any $\psi_0 \in H^s(\mathbf{R}^3)$ and any $M > C_1||\psi_0||_{H^s(\mathbf{R}^3)}$, there exists $T > 0$ depending only on $||\psi_0||_{H^s(\mathbf{R}^3)}$, s, p, γ, and M such that $\mathcal{F}$ has a unique fixed point in $\mathbf{X}(T, M)$ which is a solution of (3.1) in $L^\infty((-T, T); H^s(\mathbf{R}^3)) \cap L^\gamma((-T, T); L^\infty(\mathbf{R}^3))$.

We now call T^* the supremum of all $T > 0$ for which there exists a solution of (3.1) in $L^\infty((-T, T); H^s(\mathbf{R}^3)) \cap L^\gamma((-T, T); L^\infty(\mathbf{R}^3))$. We have proved that $T^* > 0$. Moreover, by Lemma 3.3, two such solutions coincide on the intersection of their domains of definition. We can see that if $T^* < \infty$ then no solution of (3.1) exists on $L^\infty((-T^*, T^*); H^s(\mathbf{R}^3)) \cap L^\gamma((-T^*, T^*); L^\infty(\mathbf{R}^3))$. By Lemma 3.3, if such a solution existed, it would be in $\mathbf{C}([-T^*, T^*]; H^s(\mathbf{R}^3))$. Then, however, this solution could be continued beyond T^* by solving (3.1), taking $\psi(T^*)$ as the initial value. We can then define the maximal solution ψ of (3.1) as the solution in $\mathbf{C}((-T^*, T^*); H^s(\mathbf{R}^3))$ such that for any $T \in (0, T^*)$, it belongs to $L^\infty((-T, T); H^s(\mathbf{R}^3)) \cap L^\gamma((-T, T); L^\infty(\mathbf{R}^3))$.

In order to prove the continuous dependence of the solution from the initial value, fix $T \in (0, T^*)$ and M such that (3.2) and (3.3) hold for all initial values ϕ with H^s norm less than $2||\psi_0||_{H^s(\mathbf{R}^3)}$. Then the fixed-point argument can be done in $\mathbf{X}(T, M)$ for any such ϕ. This implies that $T < T^*(\phi)$ for all of these ϕ's. Moreover, let $\psi_{0,k}$ be a sequence such that

$$||\psi_{0,k}||_{H^s(\mathbf{R}^3)} \leq 2||\psi_0||_{H^s(\mathbf{R}^3)}, \qquad ||\psi_{0,k} - \psi_0||_{H^s(\mathbf{R}^3)} \longrightarrow 0 \quad \text{as } k \longrightarrow \infty.$$

Therefore, $T < T^*(\psi_{0,k})$ and $|||\psi_k||| < M$ for all k. We now have

$$\begin{aligned} ||\psi_k - \psi||_{L^2(\mathbf{R}^3)} &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{L^2(\mathbf{R}^3)} + \int_0^t ||(G(\psi_k(\tau)) - G(\psi(\tau)))||_{L^2(\mathbf{R}^3)} d\tau \\ &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{L^2(\mathbf{R}^3)} \\ &\quad + C_p||\psi_k - \psi||_{L^\infty((0,T);L^2(\mathbf{R}^3))} \int_0^t \Big(||\psi_k(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)} + ||\psi(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)}\Big) d\tau \\ &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{L^2(\mathbf{R}^3)} + C_3 T^{1-\frac{p-1}{\gamma}} M^{p-1} ||\psi_k - \psi||_{L^2(\mathbf{R}^3)}. \end{aligned}$$

Now using (3.3), we deduce that

$$||\psi_k - \psi||_{L^2(\mathbf{R}^3)} \leq \Big(1 - C_3 T^{1-\frac{p-1}{\gamma}} M^{p-1}\Big)^{-1} ||W(t)(\psi_{0,k} - \psi_0)||_{L^2(\mathbf{R}^3)},$$

from which we get $\psi_k \longrightarrow \psi$ in $\mathbf{C}((-T, T); L^2(\mathbf{R}^3))$. Since by construction the sequence ψ_k is bounded in $\mathbf{C}((-T, T); H^s(\mathbf{R}^3))$ and, more precisely,

$$||\psi_k(t)||_{H^s} \leq M \quad \text{for } t \in (-T, T),$$

we deduce by interpolation that

$$||\psi_k(t) - \psi(t)||_{H^{s'}} \leq C(M)||W(t)(\psi_{0,k} - \psi_0)||^{\theta}_{L^2(\mathbf{R}^3)}$$

for some $\theta \in (0,1)$, from which we get $\psi_k \longrightarrow \psi$ in $\mathbf{C}((-T,T); H^{s'}(\mathbf{R}^3))$ for all $s' \in (0,s)$.

Finally, we prove that this solution ψ belongs to $\mathbf{C}^1((-T,T); L^2(\mathbf{R}^3))$. Since $\psi \in \mathbf{C}([-T,T]; H^s(\mathbf{R}^3))$, we know that $\nabla\psi \in \mathbf{C}([-T,T]; H^{s-1}(\mathbf{R}^3))$. On the other hand, we claim that G continuously maps $\mathbf{C}([-T,T]; H^s(\mathbf{R}^3))$ into $\mathbf{C}([-T,T]; L^2(\mathbf{R}^3))$. Indeed, since G continuously maps $L^{2p}(\mathbf{R}^3)$ into $L^2(\mathbf{R}^3)$, it continuously maps $\mathbf{C}([-T,T]; L^{2p}(\mathbf{R}^3))$ into $\mathbf{C}([-T,T]; L^2(\mathbf{R}^3))$. Also, it maps $\mathbf{C}([-T,T]; H^s(\mathbf{R}^3))$ into $\mathbf{C}([-T,T]; L^{2p}(\mathbf{R}^3))$ by the Sobolev embedding theorem since $1 \leq s(p) < s \Longrightarrow 2p < \frac{6}{3-2s}$. Then differentiating then the integral equation (3.1), we obtain that ψ satisfies equation (0.1), from which we get that ψ_t is in $\mathbf{C}((-T,T); L^2(\mathbf{R}^3))$ and so ψ belongs to $\mathbf{C}^1((-T,T); L^2(\mathbf{R}^3))$. □

Remark 9. Assume that the nonlinearity G is in $W^{1+s'',\infty}_{\mathrm{loc}}(\mathbf{C}^4, \mathbf{C}^4)$ for some $s'' > s$ and is such that, for some positive constant,

$$||dG||_{W^{s'',\infty}(\mathbf{B}_r)} \leq Cr^{p-1},$$

where $\mathbf{B}_r$ is the ball in $\mathbf{C}^4$ centered at zero and of radius r. Then

$$\begin{aligned} ||\psi_k - \psi||_{H^s(\mathbf{R}^3)} &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{H^s(\mathbf{R}^3)} + \int_0^t ||(G(\psi_k(\tau)) - G(\psi(\tau)))||_{H^s(\mathbf{R}^3)} d\tau \\ &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{H^s(\mathbf{R}^3)} \\ &\quad + C_p||\psi_k - \psi||_{L^\infty(0,T;H^s(\mathbf{R}^3))} \int_0^t ||dG||_{W^{s'',\infty}(\mathbf{B}_{r(k,\tau)})} d\tau, \end{aligned}$$

where $r(k,\tau) = ||\psi_k(\tau)||_{L^\infty(\mathbf{R}^3)} + ||\psi(\tau)||_{L^\infty(\mathbf{R}^3)}$. Then by Theorem 1.5,

$$\begin{aligned} ||\psi_k - \psi||_{H^s(\mathbf{R}^3)} &\leq ||W(t)(\psi_{0,k} - \psi_0)||_{H^s(\mathbf{R}^3)} \\ &\quad + C'_p||\psi_k - \psi||_{L^\infty(0,T;H^s(\mathbf{R}^3))} \int_0^t \left(||\psi_k(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)} + ||\psi(\tau)||^{p-1}_{L^\infty(\mathbf{R}^3)}\right) d\tau. \end{aligned}$$

Then it is clear that for small enough T, the application $\psi_0 \to \psi$ is Lipschitz from V to $\mathbf{C}((-T,T); H^s(\mathbf{R}^3))$. Using Theorem 1.5, one can prove in a similar way that the application is Lipschitz from V to $L^\gamma((-T,T); L^\infty(\mathbf{R}^3))$.

Proof of Theorem II. First, suppose that ψ_0 is any given function of $H^{s(p)}(\mathbf{R}^3)$. In a similar way as in the proof of Theorem I, we define

$$\mathbf{X}(T,M,\psi_0) = \{\psi \in L^\infty((-T,T); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}((-T,T); L^\infty(\mathbf{R}^3)); |||\psi||| \leq M\},$$

where

$$|||\psi||| = \sup_{t\in(-T,T)} ||\psi - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} + ||\psi||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))}.$$

By choosing T sufficiently small, we can be sure that the set $\mathbf{X}(T,M,\psi_0)$ is not empty for M as small as we need. Indeed, the function $W(t)\psi_0$ is such that

$$\lim_{t\to 0} ||W(t)\psi_0 - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} = 0, \tag{3.4}$$

and since by Theorem 1.5, $W(t)\psi_0 \in L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3))$, we also have

$$\lim_{T\to 0} ||W(t)\psi_0||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} = 0. \tag{3.5}$$

The set $\mathbf{X}(T, M, \psi_0)$ endowed with the metric

$$d(\psi_1, \psi_2) = ||\psi_1 - \psi_2||_{L^\infty((-T,T);L^2(\mathbf{R}^3))}$$

is a complete metric space.

Now there is a positive constant C_1 such that for any $\psi \in \mathbf{X}(T, M, \psi_0)$,

$$\begin{aligned} |||\mathcal{F}(\psi)||| \leq \sup_{t\in(-T,T)} ||W(t)\psi_0 - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} &+ ||W(t)\psi_0||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} \\ &+ C_1(M + ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)})M^{p-1}. \end{aligned} \tag{3.6}$$

On the other hand, by Lemma (3.2), there is a positive constant C_2 such that for every ψ_1 and ψ_2 in $\mathbf{X}(T, M, \psi_0)$, we have

$$d(\mathcal{F}(\psi_1), \mathcal{F}(\psi_2)) \leq C_2 d(\psi_1 - \psi_2)M^{p-1}. \tag{3.7}$$

Let us fix conditions on T and M in order for $\mathcal{F}$ to be a strict contraction on $\mathbf{X}(T, M, \psi_0)$. We first choose M such that

$$C_2 M^{p-1} < 1 \quad \text{and} \quad C_1(M + ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)})M^{p-2} < \frac{1}{2}. \tag{3.8}$$

Then using (3.3) and (3.4), we take T small enough in order to have $\mathbf{X}(T, M, \psi_0)$ not be empty and

$$||W(t)\psi_0||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} + \sup_{t\in(-T,T)} ||W(t)\psi_0 - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} < \frac{M}{2}. \tag{3.9}$$

Under conditions (3.8) and (3.9), $\mathcal{F}$ is a strict contraction from $\mathbf{X}(T, M, \psi_0)$ into itself. This shows that for any ψ_0 in $H^{s(p)}$, there is $T \equiv T(\psi_0)$ such that $\mathcal{F}$ has a unique fixed point in $\mathbf{X}(T, M, \psi_0)$ which is a solution of (3.1) in $L^\infty((-T, T); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}((-T, T); L^\infty(\mathbf{R}^3))$. Using Lemma 3.3 as in the proof of Theorem I, we deduce the existence of a maximal solution ψ of (3.1) in $\mathbf{C}((-T^*, T^*); H^{s(p)}(\mathbf{R}^3))$ which is unique in $L^\infty((-T, T); H^{s(p)}(\mathbf{R}^3)) \cap L^\gamma((-T, T); L^\infty(\mathbf{R}^3))$ for all $T \in (0, T^*)$.

In order to prove the continuous dependence in the same way as in Theorem I, we will show that we can find $T > 0$, $M > 0$, and a neighborhood V of ψ_0 in $H^{s(p)}(\mathbf{R}^3)$ such that the previous fixed-point argument can be done in $\mathbf{X}(T, M, \phi)$ for any $\phi \in V$. For this observe that for all $u \in \mathbf{X}(T, M, \phi)$,

$$\begin{aligned} |||\mathcal{F}(u)||| \leq & \sup_{t\in(-T,T)} ||W(t)\phi - \phi||_{H^{s(p)}(\mathbf{R}^3)} + ||W(t)\phi||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} \\ & + C(M + ||\phi||_{H^{s(p)}(\mathbf{R}^3)})M^{p-1} \\ \leq & 2||(\phi - \psi_0)||_{H^{s(p)}(\mathbf{R}^3)} + \sup_{t\in(-T,T)} ||W(t)\psi_0 - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} \\ & + ||W(t)(\phi - \psi_0)||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} \\ & + ||W(t)\psi_0||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} \\ & + C(M + ||(\phi - \psi_0)||_{H^{s(p)}(\mathbf{R}^3)} + ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)})M^{p-1}. \end{aligned} \tag{3.10}$$

On the other hand, by Lemma 3.2., for any u and v in $\mathbf{X}(T, M, \phi)$, we have (3.7), i.e.,

$$d(\mathcal{F}(u), \mathcal{F}(v)) \leq C_3 d(u, v) M^{p-1}.$$

First, we choose M such that

$$C_3 M^{p-1} < 1 \quad \text{and} \quad C(M + 2||\psi_0||_{H^{s(p)}(\mathbf{R}^3)}) M^{p-2} < \frac{1}{2}. \tag{3.11}$$

Then we choose T such that

$$\sup_{t\in(-T,T)} ||W(t)\psi_0 - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} + ||W(t)\psi_0||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} < \frac{M}{2}. \tag{3.12}$$

Finally, in order to define the neighborhood V of ψ_0 in $H^{s(p)}$, we remark that by Theorem 1.5 there is a positive constant C such that for every $T > 0$,

$$\begin{aligned} ||W(t)(\psi_0 - \phi)||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} &\leq ||W(t)(\psi_0 - \phi)||_{L^{p-1}(\mathbf{R};L^\infty(\mathbf{R}^3))} \\ &\leq C||\psi_0 - \phi||_{H^{s(p)}(\mathbf{R}^3)}. \end{aligned}$$

Therefore, there is δ_0 small enough such that if $||\phi - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} < \delta_0$, then for all $T > 0$,

$$2||\phi - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} + ||W(t)(\psi_0 - \phi)||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} < \frac{M}{4}. \tag{3.13}$$

We then define

$$V = \{\phi \in H^{s(p)}(\mathbf{R}^3); ||\phi - \psi_0||_{H^{s(p)}(\mathbf{R}^3)} < \delta\}, \tag{3.14}$$

where

$$\delta = \min(\delta_0, ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)}). \tag{3.15}$$

Then observe that if $\phi \in V$ by (3.15) and (3.11), we also have

$$\begin{aligned} C(M + ||(\phi - \psi_0)||_{H^{s(p)}(\mathbf{R}^3)} + ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)})M^{p-1} &\leq C(M + 2||\psi_0||_{H^{s(p)}(\mathbf{R}^3)})M^{p-1} \\ &< \frac{M}{4}. \end{aligned}$$

It is now straightforward to check that for every $\phi \in V$, the application $\mathcal{F}$ is a strict contraction on $\mathbf{X}(T, M, \phi)$. Therefore, for all $\phi \in V$, it has a unique fixed point in $\mathbf{X}(T, M, \phi)$ which is a solution to (3.1) in $L^\infty((-T,T); H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}((-T,T); L^\infty(\mathbf{R}^3))$. If we now consider a sequence $\{\phi_k\} \subset V$ such that $\phi_k \longrightarrow \psi_0$ in $H^{s(p)}$, then as we have just seen, $T < T^*(\phi_k)$ and $||u_k||_{L^{p-1}((-T,T);L^\infty(\mathbf{R}^3))} < M$ for all k. The proof now ends as in Theorem I.

In order to prove global existence for small initial data, consider again the estimates (3.6) and (3.7). Since we take $||\psi_0||_{H^{s(p)}(\mathbf{R}^3)} < M$, we deduce from (3.6) the existence of a positive constant C' independent of T such that

$$|||\mathcal{F}(\psi)||| \leq C'||\psi_0||_{H^{s(p)}(\mathbf{R}^3)} + 2C_1 M^p. \tag{3.16}$$

Then from (3.7) and (3.16), we see that if $M > (1 + C')||\psi_0||_{H^{s(p)}(\mathbf{R}^3)}$ and M is so small that $C'||\psi_0||_{H^{s(p)}(\mathbf{R}^3)} + 2C_1 M^p < M$ and $C_2 M^{p-1} < 1$, we may carry out our fixed-point argument in the space

$$\mathbf{X}(\infty, M, \psi_0) = \{\psi \in L^\infty(\mathbf{R}; H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3)); |||\psi||| \leq M\}.$$

We now show the scattering result. Let ψ be the global solution with small initial data ψ_0 obtained above. Since, as we have seen, $\psi \in L^\infty(\mathbf{R}; H^{s(p)}(\mathbf{R}^3)) \cap L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3))$, we deduce that $\phi_\pm$ defined by (0.6) belong to $H^{s(p)}(\mathbf{R}^3)$ and that

$$||\phi_\pm||_{H^{s(p)}(\mathbf{R}^3)} \leq ||\psi_0||_{H^{s(p)}(\mathbf{R}^3)} + C|||\psi|||^p.$$

On the other hand, for all $t \in \mathbf{R}^3$,

$$W(-t)\psi(t) - \phi_+ = \int_t^\infty W(-\tau)G(\psi(\tau))d\tau$$

from which, taking norms in $H^{s(p)}(\mathbf{R}^3)$, we get

$$\begin{aligned} ||W(-t)\psi(t) - \phi_+||_{H^{s(p)}(\mathbf{R}^3)} &\leq \int_t^\infty ||W(-\tau)G(\psi(\tau))||_{H^{s(p)}(\mathbf{R}^3)} d\tau \\ &\leq C \int_t^\infty ||\psi(\tau)||_{H^{s(p)}(\mathbf{R}^3)} ||\psi(\tau)||_\infty^{p-1} d\tau \\ &\leq C \sup_{t \in \mathbf{R}} ||u||_{H^{s(p)}(\mathbf{R}^3)} \int_t^\infty ||\psi(\tau)||_\infty^{p-1} d\tau. \end{aligned}$$

Since $\psi \in L^{p-1}(\mathbf{R}; L^\infty(\mathbf{R}^3))$, we have $||\psi(\tau)||_\infty^{p-1} \in L^1(\mathbf{R})$ and therefore

$$\lim_{t \to \infty} \int_t^\infty ||\psi(\tau)||_\infty^{p-1} d\tau = 0.$$

We then deduce

$$\lim_{t \to \infty} ||\psi(t) - W(t)\phi_+||_{H^{s(p)}(\mathbf{R}^3)} \equiv \lim_{t \to \infty} ||W(-t)\psi(t) - \phi_+||_{H^{s(p)}(\mathbf{R}^3)} = 0.$$

The same proof holds for ϕ_-. □

4. Extension to the Klein–Gordon and wave equations. The arguments of the previous sections can be used to study local and global existence for the Cauchy problem associated with the nonlinear Klein–Gordon and wave equations. For the sake of simplicity, we consider equations of the form

$$\text{(4.1)} \qquad \begin{cases} u_{tt} - \triangle u + m^2 u = (Du)^\gamma, & x \in \mathbf{R}^3, \quad t \in \mathbf{R}, \quad m \geq 0, \\ u(x,0) = f(x), & x \in \mathbf{R}^3, \\ u_t(x,0) = g(x), & x \in \mathbf{R}^3, \end{cases}$$

where $Du = (\partial_t u, \nabla u)$ and γ is a multiindex of length 4 with $|\gamma| = l \in \mathbf{Z}^+$. In order to relate our results to previous ones, for the sake of brevity, let us only mention the work of Ponce and Sideris [11]. These authors considered the more general nonlinearity $G(u, \nabla u) = u^k(\nabla u)^\gamma$, where $k \in \mathbf{Z}^+$. They showed that if $l = 2$ or $l = 3$ and $2 < s \leq \frac{5}{2}$, then for every $(f, g) \in H^s(\mathbf{R}^3) \times H^{s-1}(\mathbf{R}^3)$, there exists a $T > 0$ depending on s and $||f||_{H^s} + ||g||_{H^{s-1}}$ such that (4.1) has a unique solution u satisfying, in particular,

$$\begin{cases} u \in \bigcap_{j}^{2} \mathbf{C}^j([0,T); H^{s-j}(\mathbf{R}^3)), \\ \int_0^T ||\nabla u(t,\cdot)||_{L^\infty}^2 dt < \infty. \end{cases} \tag{4.2}$$

Moreover, if no restriction is imposed on the order of the nonlinearity in ∇u, $|\gamma| = l \in \mathbf{Z}^+$, $l \geq 3$, then the previous result holds for $s > s(l) = \frac{5}{2} - \frac{1}{l-1}$.

Using arguments similar to those used in section 3, this can be improved to obtain a global existence result for small data. In particular, we prove the following theorem.

THEOREM 4.1. *Let $|\gamma| = l \in \mathbf{Z}^+$, $l > 3$, and $s(l) = \frac{5}{2} - \frac{1}{l-1}$. Then for every $(f,g) \in H^{s(l)}(\mathbf{R}^3) \times H^{s(l)-1}(\mathbf{R}^3)$, there exists a $T^* > 0$ depending on s and of (f,g) and a solution u to* (4.1) *such that*

$$\begin{cases} u \in \mathbf{C}((-T^*,T^*); H^{s(l)}(\mathbf{R}^3)) \cap \mathbf{C}^1((-T^*,T^*); L^2(\mathbf{R}^3)), \\ \nabla u \in L^{l-1}_{\text{loc}}((-T^*,T^*); L^\infty(\mathbf{R}^3)). \end{cases}$$

Moreover,

(i) *this solution is the unique solution satisfying*

$$\forall T < T^*, \quad u \in L^\infty((-T,T); H^s(\mathbf{R}^3)) \quad \text{and} \quad \nabla u \in L^{l-1}((-T,T); L^\infty(\mathbf{R}^3));$$

(ii) *if $T^* < \infty$, then $||\nabla u||_{L^{l-1}((-T^*,T^*);L^\infty(\mathbf{R}^3))} + ||u||_{L^\infty((-T^*,T^*);H^{s(l)}(\mathbf{R}^3))} = \infty$;*

(iii) *there is $T < T^*$ and a neighborhood V of (f,g) in $H^{s(l)}(\mathbf{R}^3) \times H^{s(l)-1}(\mathbf{R}^3)$ such that for all $0 \leq s' < s(l)$, the map $(f,g) \longrightarrow u$ is continuous from V to $\mathbf{C}((-T,T); H^{s'}(\mathbf{R}^3))$; and*

(iv) *if $||f||_{H^s} + ||g||_{H^{s-1}}$ is sufficiently small, then $T^* = \infty$,*

$$u \in L^\infty(\mathbf{R}; H^{s(l)}(\mathbf{R}^3)), \quad \text{and} \quad \nabla u \in L^{l-1}(\mathbf{R}; L^\infty(\mathbf{R}^3)).$$

The proof of Theorem 4.1 is based on the following linear estimates in the same way as the proofs of Theorems I and II are based on Theorem 1.5. The new inequality which allows us to obtain global results is (4.4) below.

THEOREM 4.2. *Let $\mathcal{K}_m(t)(f,g)$ be the solution of* (1.3). *Then the following hold:*

(i) *Given $2 \leq p < \infty$, $\frac{1}{p} + \frac{1}{q} = \frac{1}{2}$, and $\alpha(p) = 2(\frac{1}{2} - \frac{1}{p})$, there is a positive constant C such that*

$$\begin{aligned} &||\mathcal{K}_m\ (\cdot)(f,g)||_{L^q(\mathbf{R};L^p(\mathbf{R}^3))} \\ &\quad \leq C\left(\left\| D_x^{\frac{\alpha(p)}{2}} (m^2 - \triangle)^{\frac{\alpha(p)}{4}} f \right\|_{L^2(\mathbf{R}^3)} + \left\| D_x^{\frac{\alpha(p)}{2}} (m^2 - \triangle)^{\frac{\alpha(p)}{4} - \frac{1}{2}} g \right\|_{L^2(\mathbf{R}^3)} \right). \end{aligned} \tag{4.3}$$

(ii) *If $p = 3$ and $s(p) = \frac{3}{2} - \frac{1}{p-1}$, then there is a positive constant C such that*

$$||\mathcal{K}_m(\cdot)(f,g)||_{L^{p-1}(\mathbf{R};L^\infty(\mathbf{R}^3))} \leq C(1+m)^{\frac{1}{l-1}}(||f||_{H^{s(p)}(\mathbf{R}^3)} + ||(m^2 - \triangle)^{-\frac{1}{2}} g||_{H^{s(p)}(\mathbf{R}^3)}). \tag{4.4}$$

(iii) *Given $s > s(p)$ and $p \geq 3$ for all $\gamma > p-1$ such that $\frac{1}{\gamma} \in (\frac{3-2s}{2}, \frac{1}{p-1})$ and all $r' > \gamma$ such that $\frac{1}{r'} \in (\frac{3-2s}{2}, \frac{1}{\gamma})$, there is a positive constant C such that for all*

$T > 0$,

$$\begin{aligned}(4.5)\qquad &||\mathcal{K}_m\ (\cdot)(f,g)||_{L^\gamma(-T,T;L^\infty(\mathbf{R}^3))}\\ &\le C(1+m)^{\frac{1}{r'}}T^{\frac{1}{\gamma}-\frac{1}{r'}}(||f||_{H^s(\mathbf{R}^3)} + ||(m^2-\triangle)^{-\frac{1}{2}}g||_{H^s(\mathbf{R}^3)}).\end{aligned}$$

Proof. The proof follows from the corresponding estimates for the operators

$$(4.6)\qquad \mathcal{K}_m^{\pm}(\cdot)(f,g) = \left(e^{\pm it\sqrt{-\triangle+m^2}}f, e^{\pm it\sqrt{-\triangle+m^2}}(-\triangle+m^2)^{-\frac{1}{2}}g\right).$$

These are obtained in a similar way as in Theorem 1.5 using Lemmas 1.2 and 1.4. □

REFERENCES

[1] M. Balabane, T. Cazenave, A. Douady, and F. Merle, *Existence of excited states for a nonlinear Dirac field*, Comm. Math. Phys., 119 (1988), pp. 153–176.

[2] P. Brenner, *On $L_p - L_q$ estimates for the wave equation*, Math. Z., 145 (1975), pp. 251–254.

[3] T. Cazenave and L. Vazquez, *Existence of localized solutions for a classical nonlinear Dirac field*, Comm. Math. Phys., 105 (1986), pp. 35–47.

[4] T. Cazenave and F. B. Weissler, *The Cauchy problem for the critical nonlinear Schrödinger equations in H^s*, Nonlinear Anal., 14 (1990), pp. 807–836.

[5] J. P. Dias and M. Figueira, *Global existence of solutions with small initial data in H^s for the massive nonlinear Dirac equations in three space dimensions*, Boll. Un. Mat. Ital. B(7), 1 (1987), pp. 861–874.

[6] M. Escobedo, *Some remarks on the density of regular mappings in Sobolev classes of S^M-valued functions*, Rev. Mat. Univ. Comput. Madrid, 1 (1988), pp. 127–144.

[7] M. J. Esteban and E. Sere, *Stationary states of the nonlinear Dirac equation: A variational approach*, Preprint 9321, CEREMADE, Université de Paris, Dauphine, Paris, 1993; Comm. Math. Phys., 171 (1995), pp. 323–350.

[8] R. Finkelstein, R. LeLevier, and M. Ruderman, *Nonlinear Spinor fields*, Phys. Rev., 83 (1951), pp. 326–332.

[9] Y. Moreau, *Existence de solutions avec petite donnée initiale dans H^2 pour une équation de Dirac non linéaire*, Portugal. Math., 46 (1989), pp. 553–565 (supplement).

[10] B. Najman, *The nonrelativistic limit of the nonlinear Dirac equation*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 9 (1992), pp. 3–12.

[11] G. Ponce and T. C. Sideris, *Local regularity of nonlinear wave equations in three space dimensions*, Comm. Partial Differential Equations, 18 (1993), pp. 169–177.

[12] A. F. Rañada, *Classical nonlinear Dirac field models of extended particles*, in Quantum Theory, Groups, Fields and Particles, D. Reidel, Amsterdam, 1982.

[13] M. Reed, *Abstract Non-Linear Wave Equations*, Lecture Notes in Math. 507, Springer-Verlag, Berlin, 1976.

[14] M. Soler, *Classical, stable nonlinear spinor field with positive rest energy*, Phys. Rev. D, 1 (1970), pp. 2766–2769.

[15] H. Triebel, Theory of Function Spaces, Birkhäuser-Verlag, Basel, Switzerland, 1983.

[16] M. Wakano, *Intensely localized solutions of the classical Dirac–Maxwell field equations*, Progr. Theoret. Phys., 35 (1966), pp. 1117–1141.

[17] F. B. Weissler, *Local existence and nonexistence for semilinear parabolic equations in L^p*, Indiana Univ. Math J., 29 (1980), pp. 79–102.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 363–380, March 1997

IMPLICIT TIME DISCRETIZATION AND GLOBAL EXISTENCE FOR A QUASI-LINEAR EVOLUTION EQUATION WITH NONCONVEX ENERGY*

G. FRIESECKE† AND G. DOLZMANN‡

Abstract. We establish global existence of weak solutions for the viscoelastic system $u_{tt} = Div(\frac{\partial\Phi}{\partial F}(Du) + Du_t)$ with nonconvex stored-energy function Φ. Unlike previous methods [P. Rybka, *Proc. Roy. Soc. Edinburgh Sect.* A, 121 (1992), pp. 101–138], our result does not require that $\frac{\partial\Phi}{\partial F}$ be globally Lipschitz continuous. Our approach is based on implicit time discretization and a compactness property of the discrete dynamical scheme not shared by energy-minimizing sequences and not known to be shared by approximation schemes of Galerkin type.

Key words. nonconvex functionals, evolution equations, implicit time discretization, viscoelasticity, solid–solid phase transitions

AMS subject classifications. 35G25, 73G25, 65N12

PII. S0036141095285958

Introduction. This article is devoted to the system of partial differential equations (PDEs)

$$u_{tt} = \text{Div}\ (\sigma(Du) + Du_t), \tag{0.1}$$

where u is a mapping from an open bounded domain $\Omega \subset \mathbb{R}^n$ to $\mathbb{R}^N$ satisfying appropriate initial and boundary conditions, and $\sigma(F) = \partial\Phi(F)/\partial F$ for some "stored-energy function" $\Phi : M^{N\times n} \to \mathbb{R}$.

To establish global existence results has required considerable effort even in the case where $n = N = 1$ [GMM, Da, A, Y, AB, P]. Previous results in higher dimensions were obtained notably by Gajewski, Gröger, and Zacharias [GGZ], Clement [C], Friedman and Nečas [FN], Engler [E1], and Rybka [R]. For work on related models involving, e.g., higher gradients, thermal effects, or memory terms, see [AHNS, BBN1, BBN2, BFS, BHJPS, BN, CH, E2, FM, HZ, KH, MNS, NR, NS, SS]. The techniques employed for (0.1) in higher dimensions (Galerkin methods [GGZ, C]; Galerkin methods combined with regularity theorems on Du [FN]; semigroup methods [E1, R]; and semigroup methods combined with regularity theorems on Du [Pec]) make the standard assumption that $\Phi(F) \sim |F|^p$ $(p \geq 2)$ for large $|F|$ and, in addition, crucially rely on at least one of the following two hypotheses:

(i) the underlying energy function Φ is convex [GGZ, C, FN, E1, Pec], or

(ii) the nonlinearity σ is globally Lipschitz continuous [GGZ, R].

Here we present a new approach to the system (0.1) which allows one to establish global existence without hypotheses (i) or (ii), provided, e.g., the following natural

*Received by the editors May 10, 1995; accepted for publication (in revised form) November 7, 1995.

http://www.siam.org/journals/sima/28-2/28595.html

†Mathematisches Institut, Universität Freiburg, Albertstraße 23b, D-79104 Freiburg, Germany (gf@noebbi.mathematik.uni-freiburg.de). The research of this author was sponsored by Deutsche Forschungsgemeinschaft Schwerpunkt "Analysis, Ergodentheorie und effiziente Simulation dynamischer Systeme."

‡Max-Plank Institute for Mathematics in the Sciences, Inselstraße 22-26, D-04103 Leipzig, Germany (georg@mis.mpg.de).

monotonicity condition at infinity is imposed:

$$(\sigma(\bar{F}) - \sigma(F)) \cdot (\bar{F} - F) \geq 0 \quad \text{for all } |F|, |\bar{F}| \geq R \quad \text{and some } R > 0. \tag{0.2}$$

In fact, this condition has already been proposed in the case where $n = N = 1$ by Andrews and Ball [AB], who employed it to prove global existence.

Our approach was partially inspired by recent work of Kinderlehrer and Pedregal [KP] and Demoulini [De] on measure-valued solutions to evolution equations with nonconvex energy functionals and involves two main steps: (i) implicit time discretization and minimization of an associated static problem at each time step ((2.1) and (2.2) in section 2) and (ii) passing to the limit in the nonlinear term $\sigma(Du)$. The feasibility of (i) relies on a simple but powerful observation, namely that the discretized counterpart of the damping term Du_t provides convexity of the static problem despite the stored-energy function Φ not being convex (Lemma 2.1). Achieving (ii) is based on a "propagation of regularity" property of the discrete scheme which may be interesting in its own right (Proposition 3.1 in section 3): it says, roughly, that the amount of oscillations in Du at any finite time t can be controlled by the amount of oscillations present in the initial data. In particular, we show that if a sequence of initial data $u^h|_{t=0}$ converges strongly in the Sobolev space $W^{1,2}(\Omega, \mathbb{R}^N)$, then the sequence of corresponding approximate solutions u^h converges strongly in $L^2(0, T; W^{1,2}(\Omega, \mathbb{R}^N))$ as the time stepsize h tends to zero.

This controllability of oscillations created in Du until time t is a subtle feature of the dynamics. First, it becomes false if the assumption of strong convergence of the $u^h|_{t=0}$ is dropped (see Example 2.1), reflecting the fact that the system (0.1) and its discretization do not regularize Du in time. Second, it is not shared by minimizing sequences (u^j, v^j) of the underlying Lyapunov function

$$E[u, v] = \int_\Omega \left(\Phi(Du) + \frac{1}{2}\, |v|^2 \right) dx \tag{0.3}$$

(which decreases with time along solutions $(u(t), u_t(t))$ of (0.1); see (2.7) and Theorem 4.1). The lack of convexity (or polyconvexity, quasiconvexity, ...) properties of the energy Φ often enforces the creation of finer and finer oscillations in Du (see [BJ1, BJ2, CK] and after them many others) and leads to weak and not strong convergence of minimizing sequences (u^j, v^j) whose limits $(u, 0)$ in particular fail to satisfy the corresponding equilibrium equations Div $\sigma(Du) = 0$.

Example 0.1 (see [S, SH, F]). Let $n = 2$, $N = 1$, $\Omega = (0, 1) \times (0, 1)$, $\Phi = (u_x^2 - 1)^2 + (u_y)^4$, and choose the boundary condition $u(x, y)|_{\partial\Omega} = x^2/2$. One easily sees that the infimum of $I[u] = \int_\Omega \Phi(Du)\, dx$ on the Sobolev space $W^{1,4}(\Omega)$ is zero but not attained; minimizing sequences strive to simultaneously achieve $u_x(x, y) \in \{\pm 1\}$ and $u_y \equiv 0$, which is impossible as the latter together with Poincaré's inequality implies $u(x, y) \equiv x^2/2$. In fact, every minimizing sequence converges weakly and not strongly in $W^{1,4}(\Omega)$ to $x^2/2$, which is neither a minimizer nor an equilibrium state of (0.1).

Theorem 4.1 below applies in particular to this example and shows that the initial-value problem is by contrast well posed.

The interesting but more difficult issue of controllability of oscillations created as time $t \to \infty$ will not be addressed here, but see [FM] for recent progress on a related model problem in one space dimension.

The system in question is of some physical interest as a model for the dynamics of coherent solid–solid phase transitions, a connection which has in fact motivated much of the analytical [AB, P, R] and numerical [KL1, KL2, S, SH] work on (0.1). In this context, u could either be

(i) a scalar function of one spatial variable (i.e., $n = N = 1$) so that (0.1) becomes the equation of one-dimensional nonlinear viscoelasticity $u_{tt} = \sigma(u_x)_x + u_{xxt}$ [AB, P];

(ii) a scalar function of two spatial variables (i.e., $n = 2$ and $N = 1$), with u representing the out-of-plane displacement field of an antiplane shear deformation [S, SH]; or

(iii) a mapping from $\Omega \subset \mathbb{R}^3$ to $\mathbb{R}^3$ representing the deformation of a body that occupies in a reference configuration the domain Ω [KL1, KL2, R].

In suitable units, (0.1) then corresponds to the equation of balance of linear momentum $u_{tt} = \operatorname{Div} T$, where the (Piola–Kirchhoff) stress tensor T is modeled by the constitutive assumption

$$T = \sigma(Du) + Du_t. \tag{0.4}$$

As pointed out in [AB, P, R, KL1, KL2, S, SH], lack of convexity of Φ (as admitted here) is the crucial mathematical feature allowing one to model phase transitions. Our hypotheses in particular include stored-energy functions Φ whose set of minima is a finite union of rotationally invariant "potential wells" as described in [BJ1, BJ2]. We emphasize, however, that from the point of view of the continuum theory of solid–solid phase transitions more sophisticated dissipation mechanisms should be studied which should in particular meet the fundamental requirements of balance of angular momentum and dynamic frame indifference which are violated by the linear damping term in the constitutive assumption (0.4).

1. The initial-boundary-value problem. Let $\Omega \subset \mathbb{R}^n$ be an open bounded domain not required to satisfy any regularity conditions, and consider for vector-valued functions $u : \Omega \times [0, \infty) \to \mathbb{R}^N$ the initial-boundary-value problem

$$u_{tt} = \operatorname{Div}(\sigma(Du) + Du_t) \quad \text{in } \Omega \times (0, \infty), \tag{1.1a}$$

$$u = g \quad \text{on } \partial\Omega \times [0, \infty), \tag{1.1b}$$

$$u = u_0 \quad \text{in } \Omega \times \{0\}, \tag{1.1c}$$

$$u_t = v_0 \quad \text{in } \Omega \times \{0\}, \tag{1.1d}$$

where g, u_0, and v_0 are given functions, $\sigma(A) = \partial\Phi(F)/\partial F$, and the stored-energy function Φ satisfies the following hypotheses:

(H1) $\Phi \in C^2(M^{N\times n})$.

(H2) There exist $c > 0$, $C > 0$, and $p \geq 2$ such that

$$c|F|^p - C \leq \Phi(F) \leq C(|F|^p + 1), \qquad |\sigma(F)| \leq C(|F|^{p-1} + 1).$$

(H3) There exists $K > 0$ such that $(\sigma(\bar{F}) - \sigma(F)) \cdot (\bar{F} - F) \geq -K|\bar{F} - F|^2$.

LEMMA 1.1. *If* Φ *satisfies* (H1), *then* (H3) *is satisfied provided one of the following holds:*

(i) *σ is monotone, i.e.,* $(\sigma(\bar{F}) - \sigma(F)) \cdot (\bar{F} - F)) \geq 0 \, \forall F$ *and* $\bar{F}$.

(ii) *σ is globally Lipschitz continuous.*

(iii) *σ is the sum of a monotone and a globally Lipschitz continuous function.*

(iv) *σ satisfies the Andrews–Ball condition* (0.2).

Proof. The first three assertions are obvious. To prove that (0.2) implies (H3), note first that the desired inequality is clear either if both $|F|$, $|\bar F| \le 2R$ (since $\sigma|_{B(2R,0)}$ is globally Lipschitz continuous) or if both $|F|$, $|\bar F| \ge R$. Thus we may assume $|F| < R$ and $|\bar F| > 2R$. Let $F(\lambda) := \bar F - \lambda(\bar F - F)/|\bar F - F|$, then $|F(\lambda)|$ is equal to $|\bar F| > 2R$ at $\lambda = 0$ and to $|F| < R$ at $\lambda = |\bar F - F|$. Thus there exists $\lambda_0 \in (0, |\bar F - F|)$ such that $|F(\lambda_0)| = 2R$. Writing $F(\lambda_0) = F_0$ and using $\bar F - F_0 = \lambda_0(\bar F - F)/|\bar F - F|$, we calculate

$$\begin{aligned}(\sigma(\bar F) - \sigma(F)) \cdot (\bar F - F) = \frac{|\bar F - F|}{\lambda_0}&(\sigma(\bar F) - \sigma(F_0)) \cdot (\bar F - F_0)\\ &+ (\sigma(F_0) - \sigma(F)) \cdot (\bar F - F)\\ \ge{}& 0 - K_0|F_0 - F||\bar F - F| \quad (\text{since } |F_0|, |\bar F| \ge R)\\ \ge{}& -K_0|\bar F - F|^2,\end{aligned}$$

where K_0 is the Lipschitz norm of $\sigma|_{B(2R,0)}$.

2. The discrete scheme. Let $p \ge 2$ be the growth exponent from (H2), fix boundary data $g \in W^{1,p}(\Omega, \mathbb{R}^N)$, and let $\mathcal{A} := \{u \in W^{1,p}(\Omega, \mathbb{R}^N) : u - g \in W_0^{1,p}(\Omega, \mathbb{R}^N)\}$. We define approximate solutions to (1.1) by means of the following implicit time-discretization scheme (compare in particular [De]). For a fixed time stepsize $h > 0$ and initial data $u_0^h \in \mathcal{A}$, $v_0^h \in L^2(\Omega, \mathbb{R}^N)$, define inductively

$$\begin{aligned}(2.1)\qquad u^{h,-1} &:= u_0^h - hv_0^h,\\ u^{h,0} &:= u_0^h,\\ u^{h,j} &:= \text{a minimizer on } \mathcal{A} \text{ of the functional } J^{h,j}[u] \quad (j \in \mathbb{N}),\end{aligned}$$

where $\mathbb{N} = \{1, 2, 3, \dots\}$ and

(2.2)
$$J^{h,j}[u] := \int_\Omega \left(\Phi(Du) + \frac{1}{2h}|Du - Du^{h,j-1}|^2 + \frac{1}{2h^2}|u - 2u^{h,j-1} + u^{h,j-2}|^2\right) dx.$$

Note that for $j \ge 1$, the minimizers $u^{h,j}$ satisfy the Euler–Lagrange equations

$$\begin{aligned}&\int_\Omega \left[\left(\sigma(Du^{h,j}) + \frac{1}{h}(Du^{h,j} - Du^{h,j-1})\right)\right.\\ (2.3)\qquad &\qquad \left.\cdot D\zeta + \frac{1}{h^2}(u^{h,j} - 2u^{h,j-1} + u^{h,j-2}) \cdot \zeta\right] dx = 0\\ &\forall \zeta \in C_0^\infty(\Omega, \mathbb{R}^N),\end{aligned}$$

which represent a weak, time-discretized version of (1.1a). Now construct a time-dependent function $u^h : \Omega \times [0, \infty) \to \mathbb{R}^N$ by interpreting the $u^{h,j}$ as values at time jh and by interpolating linearly

$$(2.4)\quad u^h(x,t) := (j - t/h)u^{h,j-1}(x) + (t/h - (j-1))u^{h,j}(x) \quad (t \in ((j-1)h, jh]).$$

We will eventually show that if the initial data $u_0^h \to u_0$ and $v_0^h \to v_0$ in, respectively, $W^{1,p}$ and L^2, then a subsequence of the u^h converges to a weak solution of the initial-boundary-value problem (1.1). As a first step, consider the issue of well definedness of the above scheme. There is indeed something to prove, namely the following lemma.

LEMMA 2.1. *If $h > 0$, $u^{h,j-1} \in \mathcal{A}$, and $u^{h,j-2} \in L^2(\Omega, \mathbb{R}^N)$, then $J^{h,j}$ attains its infimum on $\mathcal{A}$.*

Proof. (The important thing to note here is that the term in $J^{h,j}$ involving the discretized counterpart of Du_t provides convexity of the integrand with respect to Du, despite Φ not being convex.)

By hypothesis (H2) and the fact that $\mathcal{A}$ is a weakly closed subset of $W^{1,p}(\Omega, \mathbb{R}^N)$, the assertion follows from the direct method of the calculus of variations provided $J^{j,h}$ is weakly sequentially lower semicontinuous (wslsc) on $W^{1,p}(\Omega, \mathbb{R}^N)$. Introducing the notation

$$\tilde{\Phi}(F) := \Phi(F) + \frac{1}{2h}|F|^2$$

and rewriting

$$J^{h,j}[u] = \int_\Omega \left(\tilde{\Phi}(Du) - \frac{1}{h} Du \cdot Du^{h,j-1} + \frac{1}{2h}|Du^{h,j-1}|^2 + \frac{1}{2h^2}|u - 2u^{h,j-1} + u^{h,j-2}|^2 \right),$$

it is clear that $u \mapsto J^{h,j}[u] - \int_\Omega \tilde{\Phi}(Du)\,dx$ is wslsc. Also, by (H3)

$$(D\tilde{\Phi}(\bar{F}) - D\tilde{\Phi}(F)) \cdot (\bar{F} - F) \geq \left(-K + \frac{1}{h} \right) |\bar{F} - F|^2;$$

that is to say, $\tilde{\Phi}$ is convex for $h \leq h_0 := 1/K$. Standard theorems in the calculus of variations [D, Theorem 2.6] then yield $u \mapsto \int_\Omega \tilde{\Phi}(Du)\,dx$ wslsc on $W^{1,p}(\Omega, \mathbb{R}^N)$.

In fact the above arguments only require quasi convexity of $u \mapsto \int_\Omega \tilde{\Phi}(Du)\,dx$ and hence remain valid if the damping term Δu_t in (1.1a) is replaced by a weaker (rank-one-elliptic) term whose associated quadratic form is only rank-one convex rather than convex. See section 5.3.

To proceed, we introduce the energy functional

$$E[u,v] := \int_\Omega \left(\Phi(Du) + \frac{1}{2}|v|^2 \right) dx \quad (u \in W^{1,p}(\Omega, \mathbb{R}^N), v \in L^2(\Omega, \mathbb{R}^N)) \tag{2.5}$$

and the notation

$$v^{h,j} := \frac{1}{h}(u^{h,j} - u^{h,j-1}) \quad (h > 0, j \in \mathbb{N} \cup \{0\}). \tag{2.6}$$

The next lemma is a discrete analog of the dissipation identity

$$E[u(t), u_t(t)] + \int_0^t \int_\Omega |Du_t|^2 = E[u_0, v_0] \tag{2.7}$$

for smooth solutions of (1.1).

LEMMA 2.2 (discrete energy inequality). *Let $u^{h,j}, v^{h,j}$ be as defined in* (2.1) *and* (2.6). *Given $\epsilon \in (0,1)$ there exists $h_0(\epsilon) > 0$ such that for all $h \leq h_0$*

(2.8)
$$\sup_{j \in \mathbb{N} \cup \{0\}} E[u^{h,j}, v^{h,j}] + \sum_{j=1}^\infty h \int_\Omega \left((1-\epsilon)|Dv^{h,j}|^2 + \frac{1}{2} \left| \frac{v^{h,j} - v^{h,j-1}}{h^{1/2}} \right|^2 \right) \leq E[u_0^h, v_0^h].$$

Proof. To simplify the notation, we drop all superscript h's. By (H3), for $h \leq h_0 := 2\epsilon/K$, the mapping $F \mapsto \Phi(F) + (\epsilon/h)|F - Du^j(x)|^2$ is convex for all $x \in \Omega$. Thus

$$\begin{aligned}&\Phi(\bar{F}) + \frac{\epsilon}{h}|\bar{F} - Du^j(x)|^2 - \Big(\Phi(F) + \frac{\epsilon}{h}|F - Du^j(x)|^2\Big)\\ &\quad\leq \Big(\sigma(\bar{F}) + \frac{2\epsilon}{h}(\bar{F} - Du^j(x))\Big) \cdot (\bar{F} - F)\end{aligned}$$

for all x, F, and $\bar{F}$; hence for all $j \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned}&E[u^{j+1}, v^{j+1}] - E[u^j, v^j]\\ &\quad= \int_\Omega \Big[\Phi(Du^{j+1}) + \frac{\epsilon}{h}|Du^{j+1} - Du^j|^2 - \Big(\Phi(Du^j) + \frac{\epsilon}{h}|Du^j - Du^j|^2\Big)\\ &\qquad - \frac{\epsilon}{h}|Du^{j+1} - Du^j|^2 + \frac{1}{2}(|v^{j+1}|^2 - |v^j|^2)\Big]\,dx\\ &\quad\leq \int_\Omega \Big[\Big(\sigma(Du^{j+1}) + \frac{2\epsilon}{h}(Du^{j+1} - Du^j)\Big) \cdot (Du^{j+1} - Du^j)\\ &\qquad - \frac{\epsilon}{h}|Du^{j+1} - Du^j|^2 + \frac{1}{2}(|v^{j+1}|^2 - |v^j|^2)\Big]\,dx\\ &\quad= \int_\Omega \Big[\Big(\sigma(Du^{j+1}) + \frac{1}{h}(Du^{j+1} - Du^j)\Big) \cdot (Du^{j+1} - Du^j)\\ &\qquad - \frac{1-\epsilon}{h}|Du^{j+1} - Du^j|^2 + \frac{1}{2}(|v^{j+1}|^2 - |v^j|^2)\Big]\,dx.\end{aligned}$$

Now by (H2) and the fact that $Du^j \in L^p(\Omega, M^{n\times N})$ and $\sigma(Du) \in L^{p'}(\Omega, M^{n\times N})$, where $1/p' + 1/p = 1$, $\zeta = u^{j+1} - u^j$ is admissible as a test function in the Euler–Lagrange system (2.3) for u^j. So the last expression above equals

$$\begin{aligned}&\int_\Omega \Big(-(v^{j+1} - v^j) \cdot v^{j+1} - \frac{1-\epsilon}{h}|Du^{j+1} - Du^j|^2 + \frac{1}{2}(|v^{j+1}|^2 - |v^j|^2)\Big)\,dx\\ &\quad= \int_\Omega \Big(-\frac{1-\epsilon}{h}|Du^{j+1} - Du^j|^2 - \frac{1}{2}|v^{j+1} - v^j|^2\Big)\,dx.\end{aligned}$$

Applying the above estimate successively for all $j \in \mathbb{N} \cup \{0\}$ yields the assertion.

When passing to the limit $h \to 0$, it will be important to simultaneously use two different interpolations of the $u^{h,j}$ (resp., $v^{h,j}$): the piecewise linear interpolation u^h introduced in (2.4) and a piecewise constant one taking only the values $u^{h,j}$ (and called $\tilde{u}^h$ below). While the former has the advantage of being differentiable with respect to time, it is only the latter for which the exact Euler–Lagrange system (2.3) is available for every t.

For $h > 0$ and $j \in \mathbb{N} \cup \{0\}$, we let $I_{h,j} := ((j-1)h, jh]$ and define the following functions on $\Omega \times [0, \infty)$:

$$\begin{aligned}
\tilde{u}^h(x,t) &:= u^{h,j}(x) && (t \in I_{h,j}),\\
u^h(x,t) &:= (j - \tfrac{t}{h})u^{h,j-1} + (\tfrac{t}{h} - (j-1))u^{h,j} && (t \in I_{h,j}),\\
\tilde{v}^h(x,t) &:= v^{h,j}\ (= \tfrac{1}{h}(u^{h,j} - u^{h,j-1})) && (t \in I_{h,j}),\\
v^h(x,t) &:= (j - \tfrac{t}{h})v^{h,j-1} + (\tfrac{t}{h} - (j-1))v^{h,j} && (t \in I_{h,j}),\\
\tilde{w}^h(x,t) &:= w^{h,j} (:= \tfrac{1}{h}(v^{h,j} - v^{h,j-1})) && (t \in I_{h,j}).
\end{aligned} \tag{2.9}$$

Note that

$$\partial_t u^h = \tilde{v}^h, \qquad \partial_t v^h = \tilde{w}^h \tag{2.10}$$

and that the Euler–Lagrange system (2.3) can be rewritten as

$$\int_\Omega [(\sigma(D\tilde{u}^h) + D\tilde{v}^h) \cdot D\zeta + \tilde{w}^h \cdot \zeta]\, dx = 0 \quad \forall \zeta \in C_0^\infty(\Omega, \mathbb{R}^N), \quad \forall t > 0 \tag{2.11}$$

or (integrating over time)

$$\begin{aligned}
&\int_0^T \int_\Omega [(\sigma(D\tilde{u}^h) + D\tilde{v}^h) \cdot D\zeta - v^h \cdot \zeta_t]\, dx\, dt + \int_\Omega v^h \cdot \zeta\, dx|_{t=T}\\
&\qquad - \int_\Omega v^h \cdot \zeta\, dx|_{t=0} = 0 \quad \forall T > 0,\\
&\forall \zeta \in L^1(0,T; W_0^{1,p}(\Omega, \mathbb{R}^N))\\
&\qquad \cap L^2(0,T; W^{1,2}(\Omega, \mathbb{R}^N)) \cap W^{1,1}(0,T; L^2(\Omega, \mathbb{R}^N)).
\end{aligned} \tag{2.12}$$

Notation. In the next lemma and frequently below, we will abbreviate the function spaces $X(\Omega, \mathbb{R}^N)$ and $X(\Omega, M^{N\times n})$ by X and the spaces $Y(0,T;X)$ by $Y(X)$.

LEMMA 2.3 (weak convergences from the discrete energy inequality). *Assume*

$$\sup_{h>0} E[u_0^h, v_0^h] < \infty.$$

Then after extracting suitable subsequences, we have the following convergences as $h \to 0$:

$$\begin{aligned}
u^h|_{t=0} = u_0^h &\rightharpoonup u_0 \quad \mathit{in}\, W^{1,p},\\
v^h|_{t=0} = v_0^h &\rightharpoonup v_0 \quad \mathit{in}\, L^2,
\end{aligned} \tag{2.13}$$

and for every $T > 0$,

$$\begin{aligned}
\tilde{u}^h &\overset{*}{\rightharpoonup} \tilde{u} \quad \mathit{in}\ L^\infty(W^{1,p}),\\
u^h &\overset{*}{\rightharpoonup} u \quad \mathit{in}\ L^\infty(W^{1,p}), W^{1,\infty}(L^2), W^{1,2}(W^{1,2}),\\
\tilde{v}^h &\overset{*}{\rightharpoonup} \tilde{v} \quad \mathit{in}\ L^\infty(L^2), L^2(W^{1,2}),\\
v^h &\overset{*}{\rightharpoonup} v \quad \mathit{in}\ L^\infty(L^2), L^2(W^{1,2}), W^{1,2}(W^{-1,p'}),\\
\tilde{w}^h &\overset{*}{\rightharpoonup} \tilde{w} \quad \mathit{in}\ L^2(W^{-1,p'}),\\
\sigma(D\tilde{u}^h) &\overset{*}{\rightharpoonup} \tilde{\sigma} \quad \mathit{in}\ L^\infty(L^{p'}).
\end{aligned} \tag{2.14}$$

Here $W^{-1,p'}(\Omega,\mathbb{R}^N)$ denotes the usual negative Sobolev space, i.e., the completion of $L^{p'}(\Omega,\mathbb{R}^N)$ under the norm

$$\|u\|_{W^{-1,p'}} = \sup_{\zeta\in W_0^{1,p}\setminus\{0\}} \frac{|\int_\Omega u\cdot\zeta\,dx|}{\|\zeta\|_{W^{1,p}}},$$

and we say that $u^h \overset{*}{\rightharpoonup} u$ in $L^\infty(0,T;X)$ if

$$\int_0^T \langle u^h,\zeta\rangle\,dt \to \int_0^T \langle u,\zeta\rangle\,dt \quad \forall\zeta\in L^1(0,T;X),$$

where X' denotes the dual space of X. (If X is a Banach space with separable dual X', then the dual of $L^1(0,T;X)$ is the space $L^\infty(0,T;X')$ of essentially bounded, strongly measurable functions from $(0,T)$ into X'.)

Proof. The first four assertions are immediate consequences of Lemma 2.2 and (2.10). The last statement follows from the boundedness of $D\tilde{u}^h(t)$ in L^p and (H2). Finally, to prove the fifth convergence in (2.14), note that by (2.11),

$$\|\tilde{w}^h(t)\|_{W^{-1,p'}} \le \|\sigma(D\tilde{u}^h(t))\|_{L^{p'}} + \|D\tilde{v}^h(t)\|_{L^{p'}} \quad \forall t>0.$$

Now the first term is bounded in $L^\infty(L^{p'})$ and the last term in $L^2(L^2)$, so both terms are bounded in $L^2(L^{p'})$.

The following elementary lemma shows that the weak limits of piecewise constant and piecewise linear interpolations of a function given at discrete time steps actually coincide. (For an almost identical statement, see [KP].)

LEMMA 2.4. *Let $\Omega\subset\mathbb{R}^n$ be open and let $\{w^{h,j}\}_{j\in\mathbb{N},h>0}$ be a collection of functions in $L^1_{\text{loc}}(\Omega,\mathbb{R}^N)$. Define the piecewise constant and piecewise linear interpolations $\tilde{w}^h$ and w^h (as elements of $L^1_{\text{loc}}(\Omega\times[0,\infty),\mathbb{R}^N)$ by, respectively,*

$$\tilde{w}^h(x,t) := \sum_{j=1}^\infty \chi^{h,j}(t)w^{h,j}(x),$$
$$w^h(x,t) := \sum_{j=1}^\infty \chi^{h,j}(t)\left[\left(j-\frac{t}{h}\right)w^{h,j-1}(x) + \left(\frac{t}{h}-(j-1)\right)w^{h,j}(x)\right],$$

where $\chi^{h,j} := \chi_{((j-1)h,jh)}$ denotes the characteristic function of the interval $I_{h,j} := ((j-1)h,jh]$. Suppose that

$$\begin{cases} \tilde{w}^h \rightharpoonup \tilde{w}, \\ w^h \rightharpoonup w \end{cases}$$

weakly in $L^1_{\text{loc}}(\Omega\times[0,\infty),\mathbb{R}^N)$ as $h\to 0$. Then $\tilde{w}=w$.

Proof. Clearly it is enough to check that

$$\int_0^\infty\int_\Omega w\cdot\zeta\,dx\,dt = \int_0^\infty\int_\Omega \tilde{w}\cdot\zeta\,dx\,dt \quad \forall\zeta\in C_0^\infty(\Omega\times(0,\infty),\mathbb{R}^N).$$

A simple calculation shows that for h small enough,

$$\begin{aligned}
&\int_0^\infty \int_\Omega w^h(x,t)\cdot\zeta(x,t)\,dx\,dt \\
&= \int_0^\infty \int_\Omega \sum_{j=1}^\infty \chi^{h,j}(t)\left[\left(j-\frac{t}{h}\right)w^{h,j-1}(x)+\left(\frac{t}{h}-(j-1)\right)w^{h,j}(x)\right]\cdot\zeta(x,t)\,dx\,dt \\
&= \int_0^\infty \int_\Omega \tilde{w}^h(x,t)\cdot\left(\sum_{k=1}^\infty \chi^{h,k}(t)\left[\left(j-\frac{t}{h}\right)\zeta(x,t)+\left(\frac{t}{h}-(j-1)\right)\zeta(x,t+h)\right]\right)dx\,dt.
\end{aligned}$$

Since ζ has compact support, it follows that

$$\sum_{k=1}^\infty \chi^{h,k}(t)\left[\left(j-\frac{t}{h}\right)\zeta(x,t)+\left(\frac{t}{h}-(j-1)\right)\zeta(x,t+h)\right]\to\zeta$$

uniformly on $\Omega\times[0,\infty)$, and the result is a consequence of the weak convergence of w^h and $\tilde{w}^h$.

COROLLARY 2.1. *The limits obtained in Lemma* 2.3 *satisfy* $\tilde{u}=u$ *and* $\tilde{v}=v=u_t$. *In particular,* $u\in L^\infty(W_0^{1,p})\cap W^{1,\infty}(L^2)\cap W^{1,2}(W^{1,2})\cap W^{2,2}(W^{-1,p'})$ (*so that the traces* $u|_{t=0}$ *and* $u_t|_{t=0}$ *are well defined in, respectively,* $W^{1,2}$ *and* $W^{-1,p'}$) *and*

$$\int_0^T\int_\Omega[(\tilde{\sigma}+Du_t)\cdot D\zeta-u_t\cdot\zeta_t]\,dx\,dt+\int_\Omega u_t\cdot\zeta\,dx\Big|_{t=T}-\int_\Omega u_t\cdot\zeta\,dx\Big|_{t=0}=0$$

$$\forall T>0,\qquad \forall\zeta\in L^1(W_0^{1,p})\cap L^2(W^{1,2})\cap W^{1,1}(L^2),$$

$$u|_{t=0}=u_0,$$

$$u_t|_{t=0}=v_0. \tag{2.15}$$

In order to show that u is a weak solution to (1.1), it only remains to "pass to the limit in the nonlinear term," i.e., to demonstrate that under the assumptions of Lemma 2.3, $\tilde{\sigma}=\sigma(Du)$. This is in general false.

Example 2.1. *Let* $n=N=1$, $\Phi(u_x)=(u_x^2-1)^2$, $\Omega=(0,1)$, $g(x)=x/2$, $u_0^h(x)=\int_0^x\eta(x'/h)\,dx'$, *and* $v_0^h=0$, *where* $\eta(z)=-1$ *if* $z\in(n,n+1]$ *and* $n\equiv 0$ *mod* 4, $\eta(z)=1$ *otherwise. Let* $h\in\{1/4,1/8,1/12,\ldots\}$. *Then* $u^{h,j}\equiv u_0^h$ *is a solution to the discrete scheme* (2.1) (*and its interpolations* u^h, $\tilde{u}^h$ *are solutions to the continuous problem* (1.1)). *However,* $\tilde{u}_x^h=u_x^h\equiv\eta(\cdot/h)$ *does not converge strongly in* $L^1_{\mathrm{loc}}(\Omega\times(0,T))$ *as* $h\to 0$ *and* $0=\tilde{\sigma}\neq\sigma(u_x)=\sigma(1/2)$.

3. Propagation of regularity for the discrete scheme. The fact illustrated in Example 2.1 that the limiting PDE (1.1) does not regularize Du in time was already observed in [AB, P] and is a fundamental feature of evolution equations whose underlying stored-energy functions are not convex. It implies that passage to the limit in the nonlinearity σ cannot be achieved via estimating higher spatial derivatives and appealing to compact embedding theorems—neither for the scheme (2.1) analyzed here nor for approximate solutions obtained by any other method.

Note, however, that the oscillations of u_x^h in Example 2.1 were not created by the (continuous or discretized) dynamics but were already present in the initial data.

PROPOSITION 3.1 (propagation of regularity for the deformation gradient). *Let* u_0^h *and* v_0^h *be as in Lemma* 2.3, *and let* h *be the index of the subsequence delivered* *by Lemma* 2.3 *for which the convergences* (2.13) *and* (2.14) *hold. If in addition Du_0^h converges strongly in L^2 as $h \to 0$, then both Du^h and $D\tilde{u}^h$ converge strongly in $L^2(0,T;L^2)$ as $h \to 0$ $\forall\, T > 0$. More precisely,*

$$\limsup_{h\to 0} \int_0^T \int_\Omega |Du^h - Du|^2\, dx\, dt \le \limsup_{h\to 0} \int_\Omega |Du_0^h - Du_0|^2\, dx \frac{1}{2K} e^{4KT}, \tag{3.1}$$

$$\limsup_{h\to 0} \int_0^T \int_\Omega |D\tilde{u}^h - Du|^2\, dx\, dt \le \limsup_{h\to 0} \int_\Omega |Du_0^h - Du_0|^2\, dx \frac{1}{2K} e^{4KT} \tag{3.2}$$

$\forall\, T > 0$*, with K as in* (H3).

Example 2.1 shows that the assumption that the initial data converge strongly cannot be omitted.

For the large class of stored-energy functions admitted here (see (H1)–(H3)) it seems to be unknown whether an analogue of Proposition 3.1 holds for Galerkin approximations (as used for similar systems in [BBN1]) or for approximations obtained by regularization ("viscosity method" employed for a similar system in [E2]).

Estimate (3.1) also holds for a sequence of solutions to the limit system (1.1). The proof is similar but technically simpler and is left to the interested reader.

One ingredient in the proof of Proposition 3.1 will be the following result, interesting in its own right.

Proposition 3.2 (regularization for the velocity). *Let u_0^h and v_0^h be as in Lemma* 2.3*, and let h be the index of the subsequence delivered by Lemma* 2.3 *for which the convergences* (2.13) *and* (2.14) *hold. Then* (*without assuming strong convergence for the initial data*) *v^h and $\tilde{v}^h = \partial_t u^h$ converge strongly in $L^2(0,T;L^2)$ as $h \to 0$, $\forall\, T > 0$.*

This reflects the remarkable smoothing effects of the limit system (1.1) on u_t discovered by Pego [P] in one space dimension and generalized later by Rybka [R] to several dimensions in the case of globally Lipschitz-continuous nonlinearities σ.

Before embarking upon the proof of the above results, we prepare three lemmas. The first is a well-known Aubin-type result which can, for example, be found in [L, Chapter 1, Theorem 5.1].

Lemma 3.1. *Let X_s, X, and X_w be reflexive Banach spaces such that the following inclusions hold: $X_s \hookrightarrow$ (compact) $X \hookrightarrow$ (continuous) X_w. Let p_0, $p_1 \in (1,\infty)$ and $T > 0$. Then the embedding of $L^{p_0}(0,T;X_s) \cap W^{1,p_1}(0,T;X_w)$ equipped with the norm $\|\cdot\|_{L^{p_0}(X_s)} + \|\partial_t \cdot\|_{L^{p_1}(X_w)}$ into $L^{p_0}(0,T;X)$ is compact.*

The next lemma investigates more closely the connection between the piecewise-constant and the piecewise linear interpolation of a function given at discrete time steps.

Lemma 3.2. *Let X be a Banach space and $\{w^{h,j}\}_{j\in\mathbb{N},h>0}$ a collection of elements in X. Define the piecewise constant and piecewise linear interpolations $\tilde{w}^h$ and w^h* (*as elements of $L^1_{\mathrm{loc}}([0,\infty);X)$*) *by*

$$\tilde{w}^h := w^{h,j} \quad \text{on } I_{h,j},$$

$$w^h(t) := \left(j - \frac{t}{h}\right) w^{h,j-1} + \left(\frac{t}{h} - (j-1)\right) w^{h,j} \quad \text{on } I_{h,j}$$

($j \in \mathbb{N}, I_{h,j}$ *as in Lemma* 2.4). *Assume that*

$$\sup_{j\ge 1} \|w^{h,j}\|^2 \le C_1 \tag{3.3}$$

and for some $\alpha > 0$,

$$\sum_{j=1}^{\infty} h \left\| \frac{w^{h,j} - w^{h,j-1}}{h^\alpha} \right\|^2 \leq C_2. \tag{3.4}$$

Then $\forall T > 0$ and $\forall w \in L^2(0,T;X)$ with $\sup_t \|w(t)\|^2 \leq C_1$, the following estimate holds:

$$\int_0^T \|\tilde{w}^h - w\|^2 \, dt \leq 2 \int_0^T \|w^h - w\|^2 \, dt + 4hC_1 + \frac{2}{3} h^{2\alpha} C_2.$$

Proof. For $t \in I_{h,j}$, we have by construction

$$w^h(t) - w(t) = (w^{h,j} - w(t)) - \left(j - \frac{t}{h}\right)(w^{h,j} - w^{h,j-1})$$

and thus

$$\begin{aligned} \|w^{h,j} - w(t)\|^2 &= \|(w^h(t) - w(t)) + \left(j - \frac{t}{h}\right)(w^{h,j} - w^{h,j-1})\|^2 \\ &\leq 2\|w^h(t) - w(t)\|^2 + 2\left|j - \frac{t}{h}\right|^2 \|w^{h,j} - w^{h,j-1}\|^2. \end{aligned}$$

Integrating this inequality over $I_{h,j}$ yields

$$\int_{I_{h,j}} \|w^{h,j} - w(t)\|^2 \, dt \leq 2 \int_{I_{h,j}} \|w^h(t) - w(t)\|^2 \, dt + \frac{2}{3} h \|w^{h,j} - w^{h,j-1}||^2.$$

Denoting by $J := \text{int}(\frac{T}{h})$ the largest integer less than or equal to T/h, we have $|T - Jh| \leq h$ and thus

$$\begin{aligned} \int_0^T \|\tilde{w}^h(t) - w(t)\|^2 \, dt &= \sum_{j=1}^{J} \int_{I_{h,j}} \|w^{h,j} - w(t)\|^2 \, dt + \int_{Jh}^T \|\tilde{w}^h(t) - w(t)\|^2 \, dt \\ &\leq \sum_{j=1}^{J} \left\{ 2 \int_{I_{h,j}} \|w^h(t) - w(t)\|^2 \, dt + \frac{2}{3} h \|w^{h,j} - w^{h,j-1}\|^2 \right\} + 4hC_1 \\ &\leq 2 \int_0^T \|w^h(t) - w(t)\|^2 \, dt + \frac{2}{3} h^{2\alpha} \sum_{j=1}^{\infty} h \left\| \frac{w^{h,j} - w^{h,j-1}}{h^\alpha} \right\|^2 + 4hC_1. \end{aligned}$$

This immediately implies the result.

The third (elementary and well-known) lemma will be used to estimate various "harmless" terms in the proof of Proposition 3.1.

LEMMA 3.3. *Let $\tilde{\Omega} \subseteq \mathbb{R}^m$ be open, $q \in (1, \infty)$, $\phi^h \to 0$ weakly in $L^{q'}(\tilde{\Omega})$ as $h \to 0$, and Ψ be a compact subset of $L^q(\tilde{\Omega})$. Then*

$$\sup_{\psi \in \Psi} \left| \int_\Omega \phi^h \psi \, dx \right| \to 0 \quad (h \to 0).$$

Proof of Proposition 3.2. Since $v^h \rightharpoonup v = u_t$ weakly in $L^2(W^{1,2})$ and in $W^{1,2}(W^{-1,p'})$, by Lemma 3.1, $v^h \to u_t$ strongly in $L^2(L^2)$. From this, by Lemma

3.2 with $X = L^2$, $w^{h,j} = v^{h,j}$, $w = u_t$, and $\alpha = 1/2$, we obtain $\tilde{v}^h = \partial_t u^h \to u_t$ strongly in $L^2(L^2)$. Note that hypothesis (3.4) of Lemma 3.2 is satisfied thanks to the remarkable estimate

$$\sup_{h>0} \sum_{j=1}^{\infty} h \left\| \frac{v^j - v^{j-1}}{h^{1/2}} \right\|_{L^2}^2 < \infty$$

derived in Lemma 2.2, which had not been used up until now.

Proof of Proposition 3.1. Besides (2.12), we will also need another time-integrated version of the Euler–Lagrange system (2.11) which does not require the test function ζ to be differentiable in time:

(3.5)

$$\int_0^T \int_\Omega \left[(\sigma(D\tilde{u}^h) + D\tilde{v}^h) \cdot D\zeta - \tilde{v}^h \cdot \frac{\zeta(\cdot + h) - \zeta}{h} \right] dx\, dt + \fint_{T-h}^{T} \int_\Omega \tilde{v}^h \cdot \zeta(\cdot + h)\, dx\, dt$$
$$- \fint_{-h}^{0} \int_\Omega v_0^h \cdot \zeta(\cdot + h)\, dx\, dt = 0 \quad \forall T > 0, \forall \zeta \in L^1(0, T; W_0^{1,p}).$$

Testing (3.5) with $\tilde{u}^h - u$ and (2.15) with $u^h - u$ and subtracting, we have

$$\begin{aligned}
&\int_0^t \int_\Omega [\sigma(D\tilde{u}^h) \cdot (D\tilde{u}^h - Du) - \tilde{\sigma} \cdot (Du^h - Du)]\, dx\, d\tau \\
&\quad + \int_0^t \int_\Omega [D\tilde{v}^h \cdot (D\tilde{u}^h - Du) - Du_t \cdot (Du^h - Du)]\, dx\, d\tau \\
&\quad - \int_0^t \int_\Omega \left[\tilde{v}^h \cdot \left(\tilde{v}^h(\cdot + h) - \frac{u(\cdot + h) - u}{h} \right) - u_t \cdot ((u^h)_t - u_t) \right] dx\, d\tau \\
&\quad + \int_\Omega \left[\fint_{t-h}^{1} \tilde{v}^h \cdot (\tilde{u}^h(\cdot + h) - u(\cdot + h))\, d\tau - u_t(t) \cdot (u^h(t) - u(t)) \right] dx \\
&\quad - \int_\Omega v_0^h \cdot \left[\fint_{-h}^{0} (\tilde{u}^h(\cdot + h) - u(\cdot + h))\, dt - (u_0^h - u_0) \right] dx = 0.
\end{aligned}$$

Fix $T > 0$, let $t \in (0, T)$, denote the five terms above by $T_1, \dots, T_5$, and write $\eta(h)$ for quantities tending to zero as $h \to 0$ uniformly with respect to $t \in (0, T)$.

$$\begin{aligned}
T_1 &= \int_0^t \int_\Omega [(\sigma(D\tilde{u}^h) - \sigma(Du)) \cdot (D\tilde{u}^h - Du) + \sigma(Du) \cdot\ (D\tilde{u}^h - Du) \\
&\qquad\qquad - \tilde{\sigma} \cdot (Du^h - Du)]\, dx\, d\tau \\
&\geq -K \int_0^t \int_\Omega |D\tilde{u}^h - Du|^2 - \sup_{t \in (0,T)} \left| \int_{\Omega_T} (\chi_{\Omega_t} \sigma(Du)) \cdot (D\tilde{u}^h - Du)\, dx\, d\tau \right| \\
&\quad - \sup_{t \in (0,T)} \left| \int_{\Omega_T} (\chi_{\Omega_t} \tilde{\sigma}) \cdot (Du^h - Du)\, dx\, d\tau \right| \\
&\geq -K \int_0^t \int_\Omega |D\tilde{u}^h - Du|^2 - \eta(h) \\
&\geq -2K \int_0^t \int_\Omega |Du^h - Du|^2 - \eta(h).
\end{aligned}$$

Here the first inequality follows from Lemma 1.1, the second from Lemma 3.3 (with $\tilde{\Omega} = \Omega \times (0,T) =: \Omega_T$, $p = q'$, and, e.g., $\Psi = \{\chi_{\Omega_t}\sigma(Du) : t \in [0,T]\}$, $\phi^h = D\tilde{u}^h - Du$), and the third from Lemma 3.2 (with $X = L^2$, $w^{h,j} = Du^{h,j}$, and $\alpha = 1$, noting that assumption (3.4) holds by Lemma 2.2).

$$\begin{aligned}
T_2 &= \int_0^t \int_\Omega [((Du^h)_t - Du_t) \cdot (Du^h - Du) + D\tilde{v}^h \cdot (D\tilde{u}^h - Du^h)]\, dx\, d\tau \\
&= \frac{1}{2} \int_\Omega |Du^h(t) - Du(t)|^2\, dx - \frac{1}{2} \int_\Omega |Du^h(0) - Du(0)|^2\, dx \\
&\quad + \sum_{j=1}^{\infty} \int_{I_{h,j} \cap (0,t)} \int_\Omega h(j - t/h) |Dv^{h,j}|^2 \\
&\le \frac{1}{2} \int_\Omega |Du^h(t) - Du(t)|^2\, dx - \frac{1}{2} \int_\Omega |Du^h(0) - Du(0)|^2\, dx + \frac{1}{2} h^2 \sum_{j=1}^{\infty} \int_\Omega |Dv^{h,j}|^2 \\
&= \frac{1}{2} \int_\Omega |Du^h(t) - Du(t)|^2\, dx - \frac{1}{2} \int_\Omega |Du^h(0) - Du(0)|^2\, dx + \eta(h),
\end{aligned}$$

the last equality being a consequence of Lemma 2.2.

$$\begin{aligned}
|T_3| &\le \|\tilde{v}^h\|_{L^2(L^2)} \left(\|\tilde{v}^h - u_t\|_{L^2(0,T+h;L^2)} + \left\| u_t - \frac{u - u(\cdot - h)}{h} \right\|_{L^2(h,T+h;L^2)} \right) \\
&\quad + \|u_t\|_{L^2(L^2)} \|\tilde{v}^h - v_t\|_{L^2(L^2)} \\
&= \eta(h)
\end{aligned}$$

by Proposition 3.2 and the fact that $u \in W^{1,\infty}(0, 2T; L^2)$. To estimate T_4 and T_5, we will need

$$\|\tilde{u}^h - u^h\|^2_{L^\infty(0,\infty;L^2)} = \sup_{j \in \mathbb{N}} \|u^{h,j} - u^{h,j-1}\|^2_{L^2} = h^2 \sup_{j \in \mathbb{N}} \|v^{h,j}\|^2_{L^2} = \eta(h). \tag{3.6}$$

As for T_4,

$$\begin{aligned}
|T_4| &\le \|\tilde{v}^h\|_{L^\infty(-h,T;L^2)} \|\tilde{u}^h - u\|_{L^\infty(0,T+h;L^2)} + \|u_t\|_{L^\infty(L^2)} \|u^h - u\|_{L^\infty(L^2)} \\
&\le \|\tilde{v}^h\|_{L^\infty(-h,T;L^2)} \|\tilde{u}^h - u^h\|_{L^\infty(0,T+h;L^2)} \\
&\quad + (\|\tilde{v}^h\|_{L^\infty(-h,T;L^2)} + \|u_t\|_{L^\infty(L^2)}) \|u^h - u\|_{L^\infty(0,T+h;L^2)} \\
&= \eta(h)
\end{aligned}$$

by (3.6), Proposition 3.2, and the continuous embedding $W^{1,2}(L^2) \hookrightarrow L^\infty(L^2)$. Similarly,

$$|T_5| \le \|v_0^h\|_{L^2} (\|\tilde{u}^h - u\|_{L^\infty(0,h;L^2)} + \|u_0^h - u_0\|_{L^2}) = \eta(h)$$

by the weak convergence of u_0^h to u_0 in $W^{1,p}$ and the compact embedding $W^{1,p} \hookrightarrow L^2$. Collecting terms and multiplying by two,

$$\partial_t \int_0^t \!\! \int_\Omega |Du^h - Du|^2\, dx\, dt \le \int_\Omega |Du^h(0) - Du(0)|^2\, dx + \eta(h) + 4K \int_0^t \!\! \int_\Omega |Du^h - Du|^2\, dx\, d\tau$$

and thus by Gronwall's inequality,

$$\int_0^T \int_\Omega |Du^h - Du|^2\, dx\, dt \le \left(\int_\Omega |Du^h(0) - Du(0)|^2\, dx + \eta(h) \right) \frac{1}{4K} e^{4KT}.$$

Letting $h \to 0$ gives (3.1), and (3.2) follows by appealing again to Lemma 3.2 (with $X = L^2$, $w^{h,j} = Du^{h,j}$, and $\alpha = 1$). The proof of Proposition 3.1 is complete.

4. Global existence of weak solutions and the energy inequality. We are now ready to present the main result of this article.

THEOREM 4.1 (global existence of weak solutions). *Let $\Omega \subset \mathbb{R}^n$ be open and bounded, assume Φ satisfies* (H1), (H2), *and* (H3)*; let $g \in W^{1,p}(\Omega, \mathbb{R}^N)$, and let the initial data $u_0 \in \mathcal{A} = \{u \in W^{1,p}(\Omega, \mathbb{R}^N) : u - g \in W_0^{1,p}(\Omega, \mathbb{R}^N)\}$ and $v_0 \in L^2(\Omega, \mathbb{R}^N)$. Then there exists*

$$\begin{aligned} u \in\ & L^\infty(0, \infty; \mathcal{A}) \\ & \cap W^{1,\infty}(0, \infty; L^2(\Omega, \mathbb{R}^N)) \\ & \cap W^{1,2}_{\mathrm{loc}}([0, \infty); W^{1,2}(\Omega, \mathbb{R}^N)) \\ & \cap W^{2,2}_{\mathrm{loc}}([0, \infty); W^{-1,p'}(\Omega, \mathbb{R}^N)), \end{aligned} \tag{4.1}$$

which is a weak solution of (1.1)*, i.e.,*

$$\begin{aligned} \int_0^\infty \int_\Omega [(\sigma(Du) + Du_t) \cdot D\zeta - u_t \cdot \zeta_t]\, dx\, dt = 0 \quad & \forall \zeta \in C_0^\infty(\Omega \times (0, \infty), \mathbb{R}^N) \\ u|_{t=0} = u_0&, \\ u_t|_{t=0} = v_0&, \end{aligned} \tag{4.2}$$

and satisfies the dissipation inequality

$$E[u(t), u_t(t)] - E[u_0, v_0] \le -\int_0^t \int_\Omega |Du_t|^2\, dx\, d\tau \tag{4.3}$$

for almost every $t > 0$.

Proof. Set $u_0^h \equiv v_0$, $v_0^h \equiv v_0\, \forall h$. Define $u^{h,j}$ ($j \in \mathbb{N}, h > 0$) by (2.1). Then the hypotheses of Lemma 2.4 and Proposition 3.1 are trivially satisfied, and all assertions except the energy inequality follow immediately from Lemma 2.3, Corollary 2.1, and Proposition 3.1.

The energy inequality is less immediate as the mapping

$$u \mapsto \int_\Omega \Phi(Du)\, dx$$

is not assumed to be weakly lower semicontinuous on $W^{1,p}$. The remedy here is to split Φ into $(\Phi(\cdot) + \frac{K}{2}|\cdot|^2)$ and $(-\frac{K}{2}|\cdot|^2)$ (K as in (H3)) and to use the strong convergence of $D\tilde{u}^h$ in L^2.

By Lemma 2.2, $\forall \epsilon \in (0, 1)$ there exists $h_0(\epsilon) > 0$ such that for $0 < h \le h_0(\epsilon)$ and every $t > 0$, the following inequality holds:

$$\begin{aligned} & \int_\Omega \Phi(D\tilde{u}^h)\, dx + \frac{1}{2} \int_\Omega |\tilde{v}^h(t)|^2\, dx + (1 - \epsilon) \int_0^t \int_\Omega |D\tilde{v}^h|^2\, dx\, d\tau \\ & \le \int_\Omega \Phi(Du_0)\, dx + \frac{1}{2} \int_\Omega |v_0|^2\, dx. \end{aligned}$$

Since $\tilde{v}^h(t) \to u_t$ strongly in $L^2(\Omega)$ for a.e. $t > 0$ and $D\tilde{v}^h \rightharpoonup Du_t$ weakly in $L^2(\Omega_t)$, it follows from the weak lower semicontinuity of the norm in $L^2(\Omega_t)$ that

$$\begin{aligned}(4.4)\qquad \limsup_{h\to 0} \int_\Omega \Phi(D\tilde{u}^h)\,dx + \frac{1}{2}\int_\Omega |u_t|^2\,dx + (1-\epsilon)\int_0^t \int_\Omega |Du_t|^2\,dx\,d\tau \\ \leq \int_\Omega \Phi(Du_0)\,dx + \frac{1}{2}\int_\Omega |v_0|^2\,dx \quad (\text{a.e. } t > 0).\end{aligned}$$

However, since for a.e. $t > 0$

$$\begin{cases} D\tilde{u}^h(t) \rightharpoonup Du(t)\,\text{weakly in } L^p, \\ D\tilde{u}^h(t) \to Du(t)\,\text{strongly in } L^2, \end{cases}$$

we obtain

$$\begin{aligned}\int_\Omega \Phi(Du)\,dx &= \int_\Omega \left(\Phi(Du) + \frac{K}{2}|Du|^2\right)\,dx - \int_\Omega \frac{K}{2}|Du|^2\,dx \\ &\leq \limsup_{h\to 0}\int_\Omega \left(\Phi(D\tilde{u}^h) + \frac{K}{2}|D\tilde{u}^h|^2\right)\,dx - \int_\Omega \frac{K}{2}|Du|^2\,dx \\ &\leq \limsup_{h\to 0}\int_\Omega \Phi(D\tilde{u}^h)\,dx + \limsup_{h\to 0}\int_\Omega \frac{K}{2}|Du^h|^2\,dx - \int_\Omega \frac{K}{2}|Du|^2\,dx \\ &= \limsup_{h\to 0}\int_\Omega \Phi(D\tilde{u}^h)\,dx.\end{aligned}$$

Substituting this inequality into (4.4) yields the result since $\epsilon > 0$ was arbitrary.

5. Concluding remarks.

5.1. Uniqueness in the Lipschitz continuous case. If in addition σ is globally Lipschitz continuous, weak solutions of (1.1) are unique in the function class in which we have established existence. Indeed, a simple calculation shows that if u and $\bar{u}$ are two solutions of (4.2) and lie in the space (4.1), then

$$\begin{aligned}&\partial_t \frac{1}{2}\left(\int_0^t \int_\Omega |D\bar{u} - Du|^2\,dx\,d\tau + \int_\Omega |\bar{u}(t) - u(t)|^2\,dx\right) \\ &\quad = -\int_0^t \int_\Omega (\sigma(D\bar{u}) - \sigma(Du)) \cdot (D\bar{u} - Du)\,dx\,d\tau + \int_0^t \int_\Omega |\bar{u}_t - u_t|^2\,dx\,d\tau \\ &\quad \leq (\text{Lip } \sigma)\int_0^t \int_\Omega (|D\bar{u} - Du|^2 + |\bar{u}_t - u_t|^2)\,dx\,d\tau\end{aligned}$$

and

$$\begin{aligned}&\partial_t \frac{1}{2}\int_0^t \int_\Omega |\bar{u}_t - u_t|^2\,dx\,d\tau \\ &\quad = -\int_0^t \int_\Omega [(D\bar{u}_t - Du_t)\cdot(\sigma(D\bar{u}) - \sigma(Du)) + |D\bar{u}_t - Du_t|^2]\,dx\,d\tau \\ &\quad \leq \frac{1}{4}(\text{Lip } \sigma)\int_0^t \int_\Omega |D\bar{u} - Du|^2\,dx\,d\tau.\end{aligned}$$

Adding these two inequalities and appealing to Gronwall's inequality gives $\bar{u} - u \equiv 0$. This calculation is a modest generalization of the uniqueness theorem of Rybka [R] (which applies to the same assumptions on σ and to a slightly more "regular" function space).

It would be interesting to investigate whether uniqueness remains true in the more general setting of hypotheses (H1)–(H3).

5.2. Anisotropic growth at infinity. It is easily seen that Theorem 4.1 remains true if hypothesis (H2) on Φ is replaced by the anisotropic growth condition

$$\text{(H2)}' \qquad \Phi(F) = \sum_{i=1}^{N} \Phi_i(F_i), \quad \text{where } F_i \text{ denotes the } i\text{th column of } F, \quad \text{and}$$
$$\exists C > 0, \quad c > 0, \quad p_i \geq 2 \quad \text{such that for all } i,$$
$$c|f|^{p_i} - C \leq \Phi_i(f) \leq C(|f|^{p_i} + 1)$$

and $p = \max\{p_i, \dots, p_N\}$. While of minor conceptual interest, this modification entails to the best of our knowledge the first existence proof for (1.1) in the paradigm case $n = 2$, $N = 1$, $\Phi(u_x, u_y) = (u_x^2 - 1)^2 + u_y^2$, $g = 0$ studied numerically in [S, SH].

5.3. Infinitesimal frame indifference. The following modification is mathematically a little less trivial and physically more interesting. Theorem 4.1 also remains true if the system (1.1a) is replaced by the system with weaker damping

$$u_{tt} = \text{Div } \sigma(Du) + Lu_t,$$

where L is any symmetric (rank-one-) elliptic operator in divergence form, i.e., $Lv =$ Div $A(Dv)$ with A a linear mapping from $M^{N \times n}$ to $M^{N \times n}$, $A(F) \cdot G = F \cdot A(G) \, \forall F$, $G \in M^{N \times n}$,

$$A(\xi \otimes \eta) \cdot \xi \otimes \eta \geq \lambda |\xi|^2 |\eta|^2 \quad \forall \xi \in \mathbb{R}^N \text{ and } \eta \in \mathbb{R}^n \text{ and some } \lambda > 0 \tag{5.1}$$

and with the dissipation integrand $|Du_t|^2$ in (4.3) replaced by $A(Du_t) \cdot Du_t$. (To obtain approximate solutions, instead of (2.2), one would now minimize the functional

$$J^{h,j}[u] = \int_\Omega \Bigg(\Phi(Du) + \frac{1}{2h} A(Du - Du^{h,j-1}) \cdot (Du - Du^{h,j-1}) + \frac{1}{2h^2} |u - 2u^{h,j-1} + u^{h,j-2}|^2 \Bigg) \, dx,$$

whose integrand need no longer be convex but is readily shown to be quasi convex for $h \leq \lambda / K$ thanks to Plancherel's formula, (H3), and (5.1).)

In particular, in the case $n = N = 3$ relevant for elasticity, one may take $A(F) = \frac{1}{2}(F + F^T)$, i.e., $Lu_t = \text{Div}(\frac{1}{2}(Du_t + (Du_t)^T))$. This weaker damping term no longer causes energy dissipation along infinitesimal rigid rotations $u(x,t) = u_0(x) + tBx$ $(B + B^T = 0)$ and (unlike Δu_t) fulfills the fundamental requirement of frame indifference at least infinitesimally.

Acknowledgments. We thank Stefan Müller for stimulating discussions. Also, the first author thanks Hamid Bellout for interesting comments on references [BBN1, BBN2].

REFERENCES

[A] G. Andrews, *On the existence of solutions to the equation $u_{tt} = u_{xxt} + \sigma(u_x)_x$*, J. Differential Equations, 35 (1980), pp. 200–231.

[AB] G. Andrews and J. M. Ball, *Asymptotic behaviour and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[AHNS] H. W. Alt, K.-H. Hoffmann, M. Niezgódka, and J. Sprekels, *A numerical study of structural phase transitions in shape memory alloys*, preprint, Institut für Mathematik, Universität Augsburg, Augsburg, Germany, 1985.

[BBN1] H. Bellout, F. Bloom, and J. Nečas, *Existence of global weak solutions to the dynamical problem for a three-dimensional elastic body with singular memory*, SIAM J. Math. Anal., 24 (1993), pp. 36–45.

[BBN2] H. Bellout, F. Bloom, and J. Nečas, *Solutions for incompressible non-Newtonian fluids*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 795–800.

[BFS] D. Brandon, I. Fonseca, and P. J. Swart, *The creation and propagation of oscillations in a dynamical model of displacive phase transformations*, to appear.

[BHJPS] J. M. Ball, P. J. Holmes, R. D. James, R. L. Pego, and P. J. Swart, *On the dynamics of fine structure*, J. Nonlinear Sci., 1 (1991), pp. 17–70.

[BJ1] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[BJ2] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructure and the two-well problem*, Philos. Trans. Roy. Soc. London, Ser. A, 338 (1992), pp. 389–450.

[BN] H. Bellout and J. Nečas, *Existence of global weak solutions for a class of quasilinear hyperbolic integro-differential equations describing visco-elastic materials*, Math. Ann., 299 (1994), pp. 275–291.

[C] J. Clement, *Existence theorems for a quasilinear evolution equation*, SIAM J. Appl. Math., 26 (1974), pp. 745–752.

[CH] Z. Chen and K.-H. Hoffmann, *On a one-dimensional nonlinear thermoviscoelastic model for structural phase transitions in shape memory alloys*, J. Differential Equations, 112 (1994), pp. 325–350.

[CK] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal., 103 (1988), pp. 237–277.

[D] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[Da] C. M. Dafermos, *The mixed initial-boundary value problem for the equations of non-linear one dimensional viscoelasticity*, J. Differential Equations, 6 (1969), pp. 71–86.

[De] S. Demoulini, *Young measure solutions for nonlinear systems of evolution*, preprint, 1994.

[E1] H. Engler, *Global regular solutions for the dynamic antiplane shear problem in nonlinear viscoelasticity*, Math. Z., 202 (1989), pp. 251–259.

[E2] H. Engler, *Weak solutions of a class of quasilinear hyperbolic integro-differential equations describing viscoelastic materials*, Arch. Rational Mech. Anal., 113 (1991), pp. 1–38.

[F] G. Friesecke, *A necessary and sufficient condition for nonattainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 437–471.

[FM] G. Friesecke and J. B. McLeod, *Dynamics as a mechanism preventing the formation of finer and finer microstructure*, Arch. Rational Mech. Anal., 133 (1996), pp. 199–247.

[FN] A. Friedman and J. Nečas, *Systems of nonlinear wave equations with nonlinear viscosity*, Pacific J. Math., 132 (1988), pp. 29–55.

[GGZ] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatoryleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[GMM] J. M. Greenberg, R. C. MacCamy, and V. J. Mizel, *On the existence, uniqueness and stability of solutions of the equation $\sigma'(u_x)u_{xx} + \lambda u_{xtx} = \rho u_{tt}$*, J. Math. Mech., 17 (1968), pp. 707–728.

[HZ] K.-H. Hoffmann and A. Zochowski, *Analysis of the thermoelastic model of a plate with non-linear shape memory reinforcements*, Math. Methods Appl. Sci., 15 (1992), pp. 631–645.

[KH] W. D. Kalies and P. J. Holmes, *On a dynamical model for phase transformation in* *nonlinear elasticity*, preprint, 1993.

[KL] P. Klouček and M. Luskin, *Computational modeling of the martensitic transformation with surface energy*, Math. Comput. Modelling, to appear.

[KL2] P. Klouček and M. Luskin, *The computation of the dynamics of the martensitic transformation*, Contin. Mech. Thermodyn., to appear.

[KP] D. Kinderlehrer and P. Pedregal, *Weak convergence of integrands and the Young measure representation*, SIAM J. Math. Anal., 22 (1992), pp. 1–19.

[L] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[MNS] J. Milota, J. Nečas, and V. Šverák, *On weak solutions to a viscoelastic model*, Comment. Math. Univ. Carolin., 31 (1990), pp. 557–565.

[NR] J. Nečas and M. Ružička, *Global solutions to the incompressible viscous-multipolar material*, 1989, preprint.

[NS] M. Niezgodka and J. Sprekels, *Existence of solutions of a mathematical model of structural phase transitions in shape memory alloys*, Math. Methods Appl. Sci., 10 (1988), pp. 197–223.

[P] R. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Rational Mech. Anal., 97 (1987), pp. 353–394.

[Pec] H. Pecher, *On global regular solutions of third order partial differential equations*, J. Math. Anal. Appl., 73 (1980), pp. 278–299.

[R] P. Rybka, *Dynamical modeling of phase transitions by means of viscoelasticity in many dimensions*, Proc. Roy. Soc. Edinburgh Sect. A, 121 (1992), pp. 101–138.

[S] P. Swart, *The dynamical creation of microstructure in material phase transitions*, Dissertation, Cornell University, Ithaca, NY, 1991.

[SH] P. J. Swart and P. J. Holmes, *Energy minimization and the formation of microstructure in dynamic anti-plane shear*, Arch. Rational Mech. Anal., 121 (1992), pp. 37–85.

[SS] J. Sprekels and Z. Songmu, *Global solutions to the equations of a Ginzburg–Landau theory for structural phase transitions in shape memory alloys*, Phys. D., 39 (1989), pp. 59–76.

[Y] Y. Yamada, *Some remarks on the equation $y_{tt} - \sigma(y_x)y_{xx} - y_{xtx} = f$*, Osaka J. Math., 17 (1980), pp. 303–323.

SIAM J. Math. Anal.
Vol. 28, No. 2, pp. 381–388, March 1997

STRUCTURAL STABILITY OF MORSE–SMALE GRADIENT-LIKE FLOWS UNDER DISCRETIZATIONS*

MING-CHIA LI[†]

Abstract. In this paper, we show that the qualitative property of a Morse–Smale gradient-like flow is preserved by its discretization mapping obtained via numerical methods. This means that for all sufficiently small h, there is a homeomorphism H_h conjugating the time-h map Φ^h of the flow to the discretization mapping ϕ^h. Garay [*Numer. Math.*, 72 (1996), pp. 449–479] showed this result by relying on techniques of Robbin [*Ann. Math.*, 94 (1971), pp. 447–493]. Our result sharpens and unifies that in [*Numer. Math.*, 72 (1996), pp. 449–479] by using Robinson's method in [*J. Differential Equations*, 22 (1976), pp. 28–73] of the structural stability theorem for diffeomorphisms.

We also study the problem on a manifold with boundary. Under the assumption that the manifold M is positively invariant for the flow, we show that the qualitative properties are weakly stable, which means we allow the homeomorphism H_h from M into a larger manifold M' which contains M and is of the same dimension as M.

Key words. structural stability, Morse–Smale flow, gradient-like flow, numerical method

AMS subject classifications. 58F09, 58F10, 34D30, 65L06, 65L20

PII. S003614109529238X

1. Introduction.

First, we introduce some notations and definitions. Let M be a smooth manifold with distance $d(\cdot,\cdot)$ arising from the Riemannian metric. For $p \geq 0$, let $\mathrm{Diff}^p(M)$ denote the set of C^p diffeomorphisms on M with the uniformly C^p topology.

Definition 1.1. *We say that Φ^t is a C^{p+1}* Morse–Smale gradient-like flow on *M if Φ^t is a C^{p+1} flow on M and satisfies the following:* (i) *The nonwandering set $\Omega(\Phi^t)$ consists of a finite number of hyperbolic singularities; in particular, there are no closed orbits.* (ii) *The stable and unstable manifolds of singularities are transversal* (*see* [5]).

Definition 1.2. *A C^{p+1} function $\phi : \mathbb{R} \times M \to M$ is called a* discretization mapping of order p for Φ^t *if there are constants $K_1 > 0$ and $h_0 > 0$ such that $d(\Phi(h,x),\phi(h,x)) \leq K_1 h^{p+1}$ for all $h \in [0,h_0]$ and $x \in M$.*

Discretizations arise in numerical analysis. Obviously, all of the conditions above are satisfied if ϕ comes from an explicit Runge–Kutta method of order p (see [1]).

We note that for every h, Φ^h is a Morse–Smale diffeomorphism and hence it is structurally stable (see [5] or [11]). Therefore, there exists a neighborhood $\mathcal{U}_h$ of Φ^h in $\mathrm{Diff}^1(M)$ such that for any $\psi \in \mathcal{U}_h$ there is a homeomorphism H_h such that $H_h \circ \psi = \Phi^h \circ H_h$. However, the neighborhood $\mathcal{U}_h$ depends on h, so we cannot guarantee $\phi^h \in \mathcal{U}_h$ for all $h \in [0,h_0]$, although we will know $d_{C^1}(\Phi^h,\phi^h) \leq K_2 h^p$ in Lemma 2.1.

Theorem A. *Let M be a smooth compact manifold without boundary, $p \geq 2$, Φ^t be a C^{p+1} Morse–Smale gradient-like flow on M, and ϕ be a discretization mapping of order p for Φ^t. Then there is a constant $K > 0$ and for all sufficiently small h, there exists a homeomorphism H_h on M such that $H_h \circ \Phi^h(x) = \phi^h \circ H_h(x)$ and $d(H_h(x),x) \leq K h^p$ for all $x \in M$.*

*Received by the editors September 25, 1995; accepted for publication (in revised form) January 8, 1996.

http://www.siam.org/journals/sima/28-2/29238.html

[†]Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208 (minka@math.nwu.edu).

Here we only need a C^{p+1} discretization mapping with $p \geq 2$ rather than C^{p+k+1}, where $p+k \geq 3$, which Garay worked with in [2].

The following corollary is an immediate result of Theorem A which gives an observation about the qualitative embedding.

Corollary. *For all sufficiently small h, the discretization mapping ϕ^h embeds in a local flow on M. In fact, if we define $\Psi_h(t,x) = H_h \circ \Phi^t \circ H_h^{-1}(x)$ for $t \in \mathbb{R}$ and $x \in M$, then for all h small enough, $\Psi_h(t,\cdot)$ is a local flow with the properties that Ψ_h is a continuous function on $\mathbb{R} \times M$ and $\Psi_h(h,x) = \phi^h(x)$.*

Theorem B. *Let M' be a smooth manifold and M be a compact subset of M' with* closure(interior(M)) $= M$. *For $p \geq 2$, let Φ^t be a C^{p+1} Morse–Smale gradient-like flow on M' such that M is positively invariant for Φ^t and the nonwandering set of Φ^t restricted to M, $\Omega(\Phi^t|_M)$, is contained in the interior of M. Let ϕ be a discretization mapping of order p for Φ^t. Then there is a constant $K > 0$ and, for all sufficiently small h, there exists a homeomorphism H_h from M to $H_h(M) \subset M'$ such that $H_h \circ \Phi^h(x) = \phi^h \circ H_h(x)$ and $d(H_h(x),x) \leq Kh^p$ for all $x \in M$. For example, take $M' = \mathbb{R}^n$ and $M = \mathbb{D}^n$, a closed disk in $\mathbb{R}^n$.*

2. Proof of Theorem A.

2. Proof of Theorem A. Throughout this section, we assume the conditions of Theorem A. Before proving the theorem, we first investigate the d_{C^1} distance between Φ^h and ϕ^h.

Lemma 2.1. *There is a constant $K_2 > 0$ such that $d_{C^1}(\Phi^h, \phi^h) \leq K_2 h^p$ for all $h \in [0, h_0]$.*

Proof. Because $d(\Phi(h,x), \phi(h,x)) \leq K_1 h^{p+1}$, $\frac{\partial^j \Phi}{\partial h^j}(0,x) = \frac{\partial^j \phi}{\partial h^j}(0,x)$ for $j = 0, 1, \ldots, p-1$. By Taylor's expansion formula with one remainder in integral form, we have

$$\Phi(h,x) - \phi(h,x) = h^p \int_0^1 \frac{(1-s)^{p-1}}{(p-1)!} \left(\frac{\partial^p \Phi}{\partial h^p}(sh,x) - \frac{\partial^p \phi}{\partial h^p}(sh,x) \right) ds.$$

Differentiating both sides with respect to x, the integrand stays uniformly bounded since $\Phi, \phi \in C^{p+1}$ and M is compact. Therefore, there is a constant $K_2 > 0$ such that $|\frac{\partial \Phi}{\partial x}(h,x) - \frac{\partial \phi}{\partial x}(h,x)| \leq K_2 h^p$ for all $h \in [0, h_0]$ and $x \in M$. □

Since $\mathrm{Diff}^1(M)$ is an open subset of $C^1(M,M)$ (see [3]), Lemma 2.1 implies that $\phi^h \in \mathrm{Diff}^1(M)$ for all h small enough.

For the proof of Theorem A, we need more definitions. For $x, y \in M$, define $d_\Phi(x,y) = \sup\{d(\Phi^s(x), \Phi^s(y)) : s \in \mathbb{R}\}$. Then $d_\Phi(x,y)$ is a metric on the manifold M. Let $\mathfrak{X}^0(M)$ be the set of continuous vector fields on M.

Definition 2.2. *We say that a vector field $v \in \mathfrak{X}^0(M)$ is* d_Φ-Lipschitz *if there is a least positive constant $\Lambda(v)$ such that $|v(x) - v(y)| \leq \Lambda(v) d_\Phi(x,y)$ for all $x, y \in M$.*

Let $\mathfrak{X}^\Phi(M)$ be the set of all d_Φ-Lipschitz vector fields on M. For $v \in \mathfrak{X}^\Phi(M)$, we define $\|v\|_\Phi = \max\{\|v\|_0, \Lambda(v)\}$. Then $(\mathfrak{X}^\Phi(M), \|v\|_\Phi)$ is a Banach space.

Definition 2.3. *A subbundle $E \subset TM$ is* d_Φ-Lipschitz *if there is a least positive constant $\Lambda(E)$ such that $|E(x) - E(y)| \leq \Lambda(E) d_\Phi(x,y)$ for all $x, y \in M$, where $|E(x) - E(y)|$ is an appropriate distance function between Euclidean spaces.*

We briefly sketch the proof as follows. Let

$$\begin{aligned} Q_h(v_x) &= T\Phi^{-h} \circ v \circ \Phi^h(x) - \exp_x^{-1} \circ \phi^{-h} \circ \exp_{\Phi^h(x)} \circ v \circ \Phi^h(x), \\ G_h(v) &= T\Phi^{-h} \circ v \circ \Phi^h - v. \end{aligned}$$

We will construct a right inverse J_h of G_h. Let

$$\Theta_h(v) = J_h Q_h(v).$$

Then we prove that Θ_h preserves vector fields of small d_Φ-Lipschitz size, is a contraction, and has a fixed point $\widetilde{v}_h$. Thus $Q_h(\widetilde{v}_h) = G_h J_h Q_h(\widetilde{v}_h) = G_h \Theta_h(\widetilde{v}_h) = G_h(\widetilde{v}_h)$ and $\exp_x^{-1} \circ \phi^{-h} \circ \exp_{\Phi^h(x)} \circ \widetilde{v}_h \circ \Phi^h(x) = \widetilde{v}_h(x)$. For all $x \in M$, define

$$H_h(x) = \exp(\widetilde{v}_h(x)).$$

Then we have $H_h \circ \Phi^h = \phi^h \circ H_h$.

We denote the nonwandering set $\Omega(\Phi^t) = \{p_1, p_2, \ldots, p_m\}$ with the order $i \le j$ if $W^u(p_i) \cap W^s(p_j) \neq \emptyset$.

Proposition 2.4 (existence of compatible stable and unstable subbundles). *There are neighborhoods U_i of p_i, $i = 1, \ldots, m$, and compatible families of stable and unstable subbundles $\{E_i^\sigma(x) \subset T_xM : x \in O(U_i)\}$, $\sigma = s, u$. That is, the following hold:*

1. (*disjointness*) *$U_i \cap U_j = \emptyset$ for $i \neq j$.*
2. (*splitting*) *$E_i^u(x) + E_i^s(x) = T_xM$ for $x \in O(U_i)$.*
3. (*extension*) *$E_i^u(p_i) = \mathbb{E}^u(p_i)$ and $E_i^s(p_i) = \mathbb{E}^s(p_i)$, where $\mathbb{E}^u(p_i) \oplus \mathbb{E}^s(p_i) = T_{p_i}M$ is the splitting for the hyperbolic singularity p_i.*
4. (*invariance*) *E_i^u and E_i^s are $T\Phi^t$-invariant.*
5. (*compatibility*) *$E_i^u(x) \supset E_j^u(x)$ and $E_i^s(x) \subset E_j^s(x)$ if $1 \le i < j$ and $x \in O^+(U_i) \cap O^-(U_j)$.*
6. (*hyperbolicity estimate*) *There is a Riemannian metric and a constant $\mu > 0$ such that $\|T\Phi^{-t} \circ v^u \circ \Phi^t\|_0 \le e^{-\mu t}\|v^u\|_0$ and $\|T\Phi^t \circ v^s \circ \Phi^{-t}\|_0 \le e^{-\mu t}\|v^s\|_0$ for $v^u \in E_i^u|_{U_i}$, $v^s \in E_i^s|_{U_i}$, and $0 \le t \le 1$.*
7. (*d_Φ-Lipschitz*) *E_i^u and E_i^s are d_Φ-Lipschitz.*

Proof. Here we give only the outline of the proof. (For details, see [7] and [8] and also compare with the treatment for diffeomorphisms in [6] and [9].) We proceed by induction. Assume that there exist compatible families of unstable subbundles $\{E_i^u(x) : x \in O(U_i)\}$ for $i = 1, \ldots, k-1$. First, we use backward induction $j = k-1, \ldots, 1$ to construct an unstable subbundle $E_k^u(x)$ for $x \in (\cup_{l=j}^{k-1} W^u(p_l)) \cap U_{D_k^s}$, which is compatible with $E_l^u(x)$ for $j \le l \le k-1$, where $U_{D_k^s}$ is a fundamental neighborhood of $W^s(p_k)$. Then by the Φ^t-invariance, we extend the unstable subbundle E_k^u over a neighborhood U_k of p_k and $O(U_k)$. □

Choose a partition of unity $\theta_1, \ldots, \theta_m$ subordinate to the cover $O(U_1), \ldots, O(U_m)$ of M, i.e., for every i, $\theta_i : M \to [0, \infty)$ is a smooth function such that $\mathrm{supp}(\theta_i) \subset O(U_i)$ and $\sum_{i=1}^m \theta_i(x) = 1$ for all $x \in M$. For $v \in \mathfrak{X}^0(M)$, we write $\theta_i v = v_i^u + v_i^s$ with $v_i^\sigma(x) \in E_i^\sigma(x)$ for $x \in O(U_i)$ and $\sigma = s, u$. Hence $\mathrm{supp}(v_i^\sigma) \subseteq \mathrm{supp}(\theta_i) \subset O(U_i)$ for $\sigma = s, u$. Define $J_h : \mathfrak{X}^0(M) \to \mathfrak{X}^0(M)$ by

$$J_h(v) = \sum_{i=1}^{m} \left(\sum_{n=1}^{\infty} T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh} - \sum_{n=0}^{\infty} T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh} \right).$$

First, we have to show that J_h is well defined.

Proposition 2.5. *There exist $C > 1$ and $\mu > 0$ such that $\|T\Phi^r \circ v_i^s \circ v^{-r}\|_0 \le Ce^{-\mu r}\|v_i^s\|_0$ and $\|T\Phi^{-r} \circ v_i^u \circ \Phi^r\|_0 \le Ce^{-\mu r}\|v_i^u\|_0$ for all $r \ge 0$ and all i. Then J_h is well defined, a continuous linear map, and a right inverse of G_h, i.e., $G_h J_h(v) = v$.*

Proof. There exists $t_1 > 0$ such that for all $x \in M$, the orbit of x lies in $\cup_{j=0}^m U_j$ except at most t_1 amount of the time. There exists $t_2 > 0$ such that for all i, $\mathrm{supp}(\theta_i) \subset \cup_{t=-t_2}^{t_2} \Phi^t(U_i)$. Let $t_0 = t_1 + t_2$. Then for all i and $x \in \mathrm{supp}(\theta_i)$, the forward orbit of x lies in $\cup_{j=i}^m U_j$ except at most t_0 amount of the time and the backward orbit of x lies in $\cup_{j=1}^i U_j$ except at most t_0 amount of the time. It follows that there exist $C > 1$ and $\mu > 0$ such that $\|T\Phi^r \circ v_i^s \circ \Phi^{-r}\|_0 \le Ce^{-\mu r}\|v_i^s\|_0$ and $\|T\Phi^{-r} \circ v_i^u \circ \Phi^r\|_0 \le Ce^{-\mu r}\|v_i^u\|_0$ for all $r \ge 0$ and all i. Therefore, the two infinite series defining J_h converge uniformly. Since $v \mapsto v_i^\sigma$ is continuous linear, $J_h : \mathfrak{X}^0(M) \to \mathfrak{X}^0(M)$ is continuous linear. □

In order to prove the contraction of Θ_h, we first estimate J_h and Q_h in the following two lemmas.

Lemma 2.6. *There are constants $K_5, K_6 > 0$ such that $\|J_h\|_0 \le K_5 h^{-1}$ and $\Lambda(J_h(v)) \le K_6 h^{-1}(\Lambda(v) + \|v\|_0)$ for all $v \in \mathfrak{X}^\Phi(M)$.*

Proof. From Proposition 2.5, we have that $\|T\Phi^r \circ v_i^s \circ \Phi^{-r}\|_0 \le Ce^{-\mu r}\|v_i^s\|_0$ and $\|T\Phi^{-r} \circ v_i^u \circ \Phi^r\|_0 \le Ce^{-\mu r}\|v_i^u\|_0$ for all $r \ge 0$ and all i. In particular, $\|T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh}\|_0 \le Ce^{-\mu nh}\|v_i^s\|_0$ and $\|T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh}\|_0 \le Ce^{-\mu nh}\|v_i^u\|_0$ for all $n \ge 0$. Therefore,

$$\begin{aligned}\|J_h\|_0 &\le \sum_{i=1}^m 2\sum_{n=0}^\infty Ce^{-\mu nh} = \sum_{i=1}^m \frac{2C}{1-e^{-\mu h}}\\ &= \sum_{i=1}^m \frac{2C}{1-(1-\mu h + O(h^2))} = \sum_{i=1}^m \frac{2C}{\mu h - O(h^2)}\\ &\le \sum_{i=1}^m \frac{4C}{\mu} h^{-1} \le K_5 h^{-1} \quad \text{for some } K_5 > 0.\end{aligned}$$

As Robinson pointed out in [7] (see also [9]), $\Lambda(T\Phi^{-r} \circ v_i^u \circ \Phi^r) \le Ce^{-\mu r}\Lambda(v_i^u) + bCre^{-\mu(r-1)}\|v_i^u\|_0$; here b is a bound on the second derivatives in local coordinates. Thus

$$\begin{aligned}\sum_{n=0}^\infty \Lambda(T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh}) &\le \sum_{n=0}^\infty \Big(Ce^{-\mu nh}\Lambda(v_i^u) + bCnhe^{-\mu(nh-1)}\|v_i^u\|_0\Big)\\ &\le \frac{C}{1-e^{-\mu h}}\Lambda(v_i^u) + \frac{bChe^\mu}{(1-e^{-\mu h})^2 e^{\mu h}}\|v_i^u\|_0\\ &\le \frac{C}{1-e^{-\mu h}}\Lambda(v_i^u) + \frac{bChe^\mu}{(e^{\frac{\mu h}{2}} - e^{-\frac{\mu h}{2}})^2}\|v_i^u\|_0\\ &\le \frac{C}{\mu h - O(h^2)}\Lambda(v_i^u) + \frac{bChe^\mu}{(\mu h + O(h^3))^2}\|v_i^u\|_0.\end{aligned}$$

Similarly,

$$\sum_{n=1}^\infty \Lambda(T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh}) \le \frac{C}{\mu h - O(h^2)}\Lambda(v_i^s) + \frac{bChe^\mu}{(\mu h + O(h^3))^2}\|v_i^s\|_0.$$

Therefore,

$$\begin{aligned}\Lambda(J_h(v)) &\le \sum_{i=0}^m \left(\sum_{n=1}^\infty \Lambda(T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh}) + \sum_{n=0}^\infty \Lambda(T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh})\right)\\ &\le K_6 h^{-1}(\Lambda(v) + \|v\|_0) \quad \text{for some } K_6 > 0. \qquad \square\end{aligned}$$

Lemma 2.7. *There exist $K_7 > 0$ and $\delta > 0$ such that for all $\|v\|_0, \|w\|_0 < \delta$,*

$$\begin{aligned}\|Q_h(0)\|_0 &\le d_{C^0}(\phi^h, \Phi^h) \quad \textit{and}\\ \|Q_h(v) - Q_h(w)\|_0 &\le \big(K_7 \max\{\|v\|_0, \|w\|_0\} + d_{C^1}(\phi^h, \Phi^h)\big)\,\|v - w\|_0.\end{aligned}$$

In particular, $\|Q_h(v)\|_0 \leq K_7\|v\|_0\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\|v\|_0 + d_{C^0}(\phi^h, \Phi^h)$.

Proof. First, $\|Q_h(0)\|_0 = \|\exp_x^{-1} \circ \phi^{-h} \circ \Phi^h(x)\|_0 = d_{C^0}(\phi^h, \Phi^h)$. Let $L_h(v_{\Phi^h(x)}) = T\Phi^{-h}(v_{\Phi^h(x)}) - \exp_x^{-1}(\phi^{-h}(\exp_{\Phi^h(x)}(v_{\Phi^h(x)})))$. Since Φ^h and ϕ^h are C^{p+1}, L_h is $C^p \subset C^2$ and so there exist $K_7 > 0$ and $\delta > 0$ such that $\|D^2L_h(v)\|_0 \leq K_7$ for all $\|v\|_0 < \delta$. By the mean-value theorem, we have for all $\|v\|_0, \|w\|_0 < \delta$,

$$\begin{aligned}
\|Q_h(v) - Q_h(w)\|_0 &= \sup_{x\in M} |L_h(v_{\Phi^h(x)}) - L_h(w_{\Phi^h(x)})| \\
&= \sup_{y\in M} |L_h(v_y) - L_h(w_y)| \\
&= \sup_{y\in M} \left| \int_0^1 DL_h(w_y + s(v_y - w_y))(v_y - w_y)ds \right| \\
&\leq \sup_{\substack{y\in M \\ |v_y^*|\leq \|v\|_0, \|w\|_0}} |DL_h(v_y^*)| \cdot |v_y - w_y| \\
&= \sup_{\substack{y\in M \\ |v_y^*|\leq \|v\|_0, \|w\|_0}} \left\{ \left| \int_0^1 D^2L_h(sv_y^*)v_y^* ds \right| + \|DL_h(0)\|_0 \right\} \|v - w\|_0 \\
&\leq \left(K_7 \max\{\|v\|_0, \|w\|_0\} + d_{C^1}(\phi^h, \Phi^h)\right) \|v - w\|_0. \qquad \square
\end{aligned}$$

Now we show that Θ_h is a contraction and has a fixed point $\widetilde{v}_h$.

PROPOSITION 2.8. *There is a positive constant K such that for all h sufficiently small, Θ_h preserves $\mathfrak{X}^0_{Kh^p}(M)$ and is a contraction on $\mathfrak{X}^0_{Kh^p}(M)$, where $\mathfrak{X}^0_{Kh^p}(M) \equiv \{v \in \mathfrak{X}^0(M) : \|v\|_0 \leq Kh^p\}$. Therefore, for all h sufficiently small, Θ_h has a unique fixed point $\widetilde{v}_h$ in $\mathfrak{X}^0_{Kh^p}(M)$.*

Proof. From the previous lemma, we have for all $\|v\|_0, \|w\|_0 < \delta$,

$$\begin{aligned}
\|\Theta_h(v) - \Theta_h(w)\|_0 &\leq \|J_h\|_0 \|Q_h(v) - Q_h(w)\|_0 \\
&\leq K_5 h^{-1} \left(K_7 \max\{\|v\|_0, \|w\|_0\} + d_{C^1}(\phi^h, \Phi^h)\right) \|v - w\|_0.
\end{aligned}$$

Choose a suitable $K > 0$ such that for all sufficiently small h and all $v, w \in \mathfrak{X}^0_{Kh^p}(M)$, $Kh^p < \delta$,

$$\begin{aligned}
\|\Theta_h(v)\|_0 &\leq \|J_h\|_0 \left(\|Q_h(v)\|_0 + \|Q_h(0)\|_0\right) \\
&\leq K_5 h^{-1} \left(K_7 \|v\|_0\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\|v\|_0 + d_{C^0}(\phi^h, \Phi^h)\right) \\
&\leq K_5 h^{-1} \left(K_7 Kh^p Kh^p + K_2 h^p Kh^p + K_1 h^{p+1}\right) \leq Kh^p
\end{aligned}$$

and

$$\|\Theta_h(v) - \Theta_h(w)\|_0 \leq K_5 h^{-1} \left(K_7 Kh^p + K_2 h^p\right) \|v - w\|_0 < \|v - w\|_0. \qquad \square$$

Thus far, we have been able to construct a topological semiconjugacy. Since $\widetilde{v}_h$ is continuous, $H_h \equiv \exp(\widetilde{v}_h)$ is continuous. Because H_h is homotopic to the identity, H_h is of degree one and hence onto (see [4]). Moreover, $d_{C^0}(H_h, id_M) = d_{C^0}(\exp(\widetilde{v}_h), id_M) = \|\widetilde{v}_h\|_0 \leq Kh^p$. Finally, we have to prove that H_h is one to one.

To this end, we estimate the d_Φ-Lipschitz sizes of $Q_h(v)$ and $\Theta_h(v)$.

LEMMA 2.9. *For $v \in \mathfrak{X}^\Phi(M)$,*

$$\begin{aligned}
\Lambda(Q_h(v)) &\leq \left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(1 + \Lambda(v)), \\
\Lambda(\Theta_h(v)) &\leq K_6 h^{-1}\{\left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(1 + \Lambda(v)) \\
&\qquad + K_7\|v\|_0\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\|v\|_0 + d_{C^0}(\phi^h, \Phi^h)\}.
\end{aligned}$$

Thus there is a constant $K_8 > 0$ such that Θ_h preserves $v \in \mathfrak{X}^0_{Kh^p}(M) \cap \mathfrak{X}^\Phi(M)$ with $\Lambda(v) \leq K_8 h^{p-1}$.

Proof.

$$\begin{aligned}|Q_h(v_x) - Q_h(v_y)| &\leq \|DL_h(v^*)\|_0 d(v \circ \Phi^{-h}(x), v \circ \Phi^{-h}(y)) \\ &\leq \left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right) \\ &\quad \cdot \left(d(\Phi^{-h}(x), \Phi^{-h}(y)) + |v \circ \Phi^{-h}(x) - v \circ \Phi^{-h}(y)|\right) \\ &\leq \left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(d_\Phi(x,y) + \Lambda(v) d_\Phi(x,y)) \\ &= \left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(1 + \Lambda(v)) d_\Phi(x,y).\end{aligned}$$

Then $\Lambda(Q_h(v)) \leq \left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(1 + \Lambda(v))$. Moreover,

$$\begin{aligned}\Lambda(\Theta_h(v)) &\leq \Lambda(J_h Q_h(v)) \\ &\leq K_6 h^{-1}(\Lambda(Q_h(v)) + \|Q_h(v)\|_0) \\ &\leq K_6 h^{-1}\{\left(K_7\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\right)(1 + \Lambda(v)) \\ &\quad + K_7\|v\|_0\|v\|_0 + d_{C^1}(\phi^h, \Phi^h)\|v\|_0 + d_{C^0}(\phi^h, \Phi^h)\}.\end{aligned}$$

Choose a suitable constant $K_8 > 0$ such that, for all sufficiently small h and all $v \in \mathfrak{X}^0_{Kh^p}(M) \cap \mathfrak{X}^\Phi(M)$ with $\Lambda(v) \leq K_8 h^{p-1}$,

$$\begin{aligned}\Lambda(\Theta_h(v)) &\leq K_6 h^{-1}\{(K_7 K h^p + K_2 h^p)(1 + K_8 h^{p-1}) \\ &\quad + K_7 K h^p K h^p + K_2 h^p K h^p + K_1 h^{p+1}\} \\ &\leq K_8 h^{p-1}. \qquad \square\end{aligned}$$

Furthermore, the d_Φ-Lipschitz size of the fixed point $\widetilde{v}_h$ is dominated by $K_8 h^{p-1}$.

LEMMA 2.10. *For the fixed point $\widetilde{v}_h$, we have $\widetilde{v}_h \in \mathfrak{X}^\Phi(M)$ and $\Lambda(\widetilde{v}_h) \leq K_8 h^{p-1}$.*

Proof. Let $v_h^1 = 0$ and $v_h^{n+1} = \Theta_h(v_h^n)$ for $n \in \mathbb{N}$. Since Θ_h preserves the space $\mathfrak{X}^0_{Kh^p}$ and is a contraction, $\|v_h^n\|_0 \leq Kh^p$ for all $n \in \mathbb{N}$ and $\|v_h^n - \widetilde{v}_h\|_0 \to 0$ as $n \to \infty$, where $\widetilde{v}_h$ is the fixed point of Θ_h in $\mathfrak{X}^0_{Kh^p}(M)$. Because $\Lambda(v_h^n) \leq K_8 h^{p-1}$ for all $n \in \mathbb{N}$, we have $\widetilde{v}_h \in \mathfrak{X}^\Phi(M)$ and $\Lambda(\widetilde{v}_h) \leq K_8 h^{p-1}$. $\square$

We also need the following two lemmas. The first one is an easy consequence of Gronwall's theorem (see [11]). The second one was proved by Robbin in [6, Lemma 2.3].

LEMMA 2.11. *There is a constant $L > 0$ such that $d(\Phi^s(p), \Phi^s(q)) \leq L d(p,q)$ whenever $p, q \in M, s \in [0, h_0]$.*

LEMMA 2.12. *There is a constant $\alpha > 0$ such that for $p, q \in M$ and $v \in \mathfrak{X}^0_{Kh^p}(M)$, $\alpha d(p,q) - d(\exp(v(p)), \exp(v(q))) \leq |v(p) - v(q)|$.*

Now let us prove that H_h is one to one and complete the proof of Theorem A.

PROPOSITION 2.13. *For all h sufficiently small, $H_h \equiv \exp(\widetilde{v}_h)$ is one to one.*

Proof. Suppose $H_h(x) = H_h(y)$. By the conjugacy, we have $H_h(\Phi^{kh}(x)) = \phi^{kh}(H_h(x)) = \phi^{kh}(H_h(y)) = H_h(\Phi^{kh}(y))$ for all $k \in \mathbb{Z}$. There exists $s_0 \in \mathbb{R}$ such that $d_\Phi(x,y) \leq 2d(\Phi^{s_0}(x), \Phi^{s_0}(y))$. Take $k_0 \in \mathbb{Z}$ with $k_0 h \leq s_0 < (k_0+1)h$; then $s_0 - k_0 h \in [0, h) \subset [0, h_0]$. Let $p = \Phi^{k_0 h}(x)$ and $q = \Phi^{k_0 h}(y)$. By Lemma 2.11, there is a constant L such that $d(\Phi^{s_0}(x), \Phi^{s_0}(y)) = d(\Phi^{s_0 - k_0 h}(p), \Phi^{s_0 - k_0 h}(q)) \leq L d(p,q)$. Hence $d_\Phi(p,q) = d_\Phi(x,y) \leq 2d(\Phi^{s_0}(x), \Phi^{s_0}(y)) \leq 2L d(p,q)$. By the previous lemma, there exists $\alpha > 0$ such that $\alpha d(p,q) - d(\exp(\widetilde{v}_h(p)), \exp(\widetilde{v}_h(q))) \leq |\widetilde{v}_h(p) - \widetilde{v}_h(q)|$.

Because $\Lambda(\widetilde{v}_h) \le K_8 h^{p-1}$,

$$\begin{aligned}\alpha d(p,q) - d(\exp(\widetilde{v}_h(p)), \exp(\widetilde{v}_h(q))) &\le |\widetilde{v}_h(p) - \widetilde{v}_h(q)| \\ &\le K_8 h^{p-1} d_\Phi(p,q) \le K_8 h^{p-1} 2L d(p,q).\end{aligned}$$

Therefore, $(\alpha - K_8 h^{p-1} 2L) d(p,q) \le d(H_h(p), H_h(q)) = 0$. If h is small enough such that $\alpha - K_8 h^{p-1} 2L > 0$, then $d(p,q) = 0$ and $p = q$. Thus $x = \Phi^{-k_0 h}(p) = \Phi^{-k_0 h}(q) = y$, and hence H_h is one to one. □

3. Proof of Theorem B. Because $\Omega(\Phi^t|_M) \subset \text{interior}(M)$, the singularities $\{p_i : i = 1, \dots, m\}$ are not on the boundary of M. For our convenience, we denote the boundary of M by p_0. We use induction to construct compatible families of stable and unstable subbundles. Assume that for $0 \le i \le k-1$ there exist a neighborhood U_i of p_i in M and continuous subbundles $\{E_i^\sigma(x) \subset T_x M : x \in O(U_i)\}$, $\sigma = u, s$, such that the following hold:

1. (disjointness) $U_i \cap U_j = \emptyset$ for $i \ne j$.
2. (splitting) $E_i^u(x) + E_i^s(x) = T_x M$ for $x \in O(U_i) \cap M$.
3. (boundary) $E_0^u(x) = T_x M$ and $E_0^s(x) = \{0_x\}$ for $x \in O(U_0) \cap M$.
4. (extension) For $i \ne 0$, $E_i^u(p_i) = \mathbb{E}^u(p_i)$ and $E_i^s(p_i) = \mathbb{E}^s(p_i)$, where $\mathbb{E}^u(p_i) \oplus \mathbb{E}^s(p_i) = T_{p_i} M$ is the splitting for the hyperbolic singularity p_i.
5. (invariance) E_i^u and E_i^s are $T\Phi^t$-invariant.
6. (compatibility) $E_i^u(x) \supset E_j^u(x)$ and $E_i^s(x) \subset E_j^s(x)$ if $0 \le i < j$ and $x \in O^+(U_i) \cap O^-(U_j) \cap M$.
7. (hyperbolicity estimate) There is a Riemannian metric and a constant $\mu > 0$ such that $\|T\Phi^{-t} \circ v^u \circ \Phi^t\|_0 \le e^{-\mu t} \|v^u\|_0$ and $\|T\Phi^t \circ v^s \circ \Phi^{-t}\|_0 \le e^{-\mu t} \|v^s\|_0$ for $v^u \in E_i^u|_{U_i}$, $v^s \in E_i^s|_{U_i}$, and $0 \le t \le 1$.
8. (d_Φ-Lipschitz) E_i^u and E_i^s are d_Φ-Lipschitz.

First, we use backward induction $j = k-1, \dots, 0$ to construct a unstable subbundle $E_k^u(x)$ for $x \in (\cup_{l=j}^{k-1} W^u(p_l)) \cap U_{D_k^s}$, which is compatible with $E_l^u(x)$ for $j \le l \le k-1$, where $U_{D_k^s}$ is a fundamental neighborhood of $W^s(p_k)$. When $j = 0$, the only requirements for the compatibility are $E_k^u(x) \subset E_0^u(x) = T_x M$ and $\{0_x\} = E_0^s(x) \subset E_k^s(x)$, which are easily satisfied. Then by the Φ^t-invariance, we extend the unstable subbundle E_k^u over a neighborhood U_k of p_k and $O(U_k)$ (for details, see [10]).

By shrinking U_i if necessary, we can choose a partition of unity $\theta_0, \dots, \theta_m$ subordinate to the cover $O(U_0), \dots, O(U_m)$ of M such that $\text{supp}(\theta_0) \subset M$ and $\text{supp}(\theta_i) \subset \text{interior}(M)$ for $i \ne 0$. For $v \in \mathfrak{X}^0(M)$, we write $\theta_i v = v_i^u + v_i^s$ with $v_i^\sigma(x) \in E_i^\sigma(x)$ and $\sigma = u, s$. Define $J_h : \mathfrak{X}^0(M) \to \mathfrak{X}^0(M)$ by $J_h(v_x) = \sum_{i=0}^m (\sum_{n=1}^\infty T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh}(x) - \sum_{n=0}^\infty T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh}(x))$. Since M is positively invariant for Φ^h, $E_0^s(x) = \{0_x\}$ for $x \in O(U_0) \cap M$, $\text{supp}(\theta_0) \subset M$, and $\text{supp}(\theta_i) \subset \text{int}(M)$ for $i \ne 0$, it follows that $J_h(v) = -\sum_{n=0}^\infty T\Phi^{-nh} \circ v_0^u \circ \Phi^{nh} + \sum_{i=1}^m (\sum_{n=1}^\infty T\Phi^{nh} \circ v_i^s \circ \Phi^{-nh} - \sum_{n=0}^\infty T\Phi^{-nh} \circ v_i^u \circ \Phi^{nh})$. By the same argument as above, we can prove that J_h is a continuous linear map and J_h is a right inverse of G_h on $\mathfrak{X}^0(M)$. Moreover, we can prove that $\Theta_h \equiv J_h Q_h$ preserves $\mathfrak{X}^0_{Kh^p}(M)$ and is a contraction on $\mathfrak{X}^0_{Kh^p}(M)$, where $\mathfrak{X}^0_{Kh^p}(M) = \{v \in \mathfrak{X}^0(M) : \|v\|_0 \le Kh^p\}$. Then there is a unique fixed point $\widetilde{v}_h \in \mathfrak{X}^0_{Kh^p}(M)$. Therefore, $H_h \equiv \exp(\widetilde{v}_h)$ is a homeomorphism from M to $H_h(M) \subset M'$.

Remark. The referee observed that Theorem B follows from Theorem A through the following argument. Find a slightly larger M and modify the vector field so that it points into M. Then form the double N and extend the vector field by the symmetry of the double. Make the vector field on the double Morse–Smale by applying the Kupka–Smale theorem on $N\backslash M$. Finally, extend the discretization to the double. However, we have presented our original proof in order to avoid discussing the double of a manifold.

Acknowledgment. The author would like to thank Professor Clark Robinson of Northwestern University for his inspiration and general introduction to this subject.

REFERENCES

[1] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, New York, 1987.

[2] B. M. Garay, *On structural stability of ordinary differential equations with respect to discretization methods*, Numer. Math., 72 (1996), pp. 449–479.

[3] M. Hirsch, *Differential Topology*, Springer-Verlag, New York, 1976.

[4] J. R. Munkres, *Elementary Differential Topology*, Ann. Math. Stud., 54, Princeton University Press, Princeton, NJ, 1963.

[5] J. Palis and S. Smale, *Structural stability theorems*, in Global Analysis, Proc. Sympos. Pure Math. 14, AMS, Providence, RI, 1970, pp. 223–231.

[6] J. Robbin, *A structural stability theorem*, Ann. Math., 94 (1971), pp. 447–493.

[7] C. Robinson, *Structural stability of vector fields*, Ann. Math., 99 (1974), pp. 154–175.

[8] C. Robinson, *Structural stability of C^1 flows*, in Lectures Notes in Math. 468, Springer-Verlag, Berlin, 1975, pp. 262–277.

[9] C. Robinson, *Structural stability of C^1 diffeomorphisms*, J. Differential Equations, 22 (1976), pp. 28–73.

[10] C. Robinson, *Structural stability on manifolds with boundary*, J. Differential Equations, 37 (1980), pp. 1–11.

[11] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics and Chaos*, CRC Press, Boca Raton, FL, 1996.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 389–401, March 1997

RAZUMIKHIN-TYPE THEOREMS ON EXPONENTIAL STABILITY OF NEUTRAL STOCHASTIC FUNCTIONAL DIFFERENTIAL EQUATIONS[*]

XUERONG MAO[†]

Abstract. Recently, we initiated in [*Systems Control Lett.*, 26 (1995), pp. 245–251] the study of exponential stability of neutral stochastic functional differential equations, and in this paper, we shall further our study in this area. We should emphasize that the main technique employed in this paper is the well-known Razumikhin argument and is completely different from those used in our previous paper [*Systems Control Lett.*, 26 (1995), pp. 245–251]. The results obtained in [*Systems Control Lett.*, 26 (1995), pp. 245–251] can only be applied to a certain class of neutral stochastic functional differential equations excluding neutral stochastic differential delay equations, but the results obtained in this paper are more general, and they especially can be used to deal with neutral stochastic differential delay equations. Moreover, in [*Systems Control Lett.*, 26 (1995), pp. 245–251], we only studied the exponential stability in mean square, but in this paper, we shall also study the almost sure exponential stability. It should be pointed out that although the results established in this paper are applicable to more general neutral-type equation, for a particular type of equation discussed in [*Systems Control Lett.*, 26 (1995), pp. 245–251], the results there are sharper.

Key words. exponential stability, Razumikhin-type theorem, Brownian motion, Doob martingale inequality, Borel–Cantelli lemma

AMS subject classifications. 60H20, 34D08, 60G48

PII. S0036141095290835

1. Introduction.

Deterministic neutral functional differential equations and their stability have been studied by many authors, e.g., Haddock et al. [3], Hale and Lunel [4], and the references therein. Motivated by the chemical-engineering systems as well as the theory of aeroelasticity, Kolmanovskii and Nosov [8] introduced the neutral stochastic functional differential equations of the form

$$d[x(t) - G(x_t)] = f(t, x_t)dt + g(t, x_t)dw(t) \tag{1.1}$$

on $t \geq 0$ with initial data $x_0 = \xi \in L^2_{\mathcal{F}_0}([-\tau, 0]; R^n)$. (For notation, please see section 2 below.) Kolmanovskii and Nosov [8] not only established the theory of existence and uniqueness of the solution to (1.1) but also investigated the stability and asymptotic stability of the equations (see also Kolmanovskii and Myshkis [7]). However, the exponential stability of such equations has not been studied until recently by the author in [11]. To be more precise, let us give the definition of exponential stability.

DEFINITION 1.1. *Denote by $x(t;\xi)$ the solution of equation* (1.1). *The trivial solution of equation* (1.1) *is said to be exponentially stable in mean square if there exists a pair of positive constants γ and M such that*

$$E|x(t;\xi)|^2 \leq Me^{-\gamma t} \sup_{-\tau \leq \theta \leq 0} E|\xi(\theta)|^2, \quad t \geq 0,$$

[*] Received by the editors August 25, 1995; accepted for publication (in revised form) December 28, 1995. This research was partially supported by the Royal Society and the London Mathematical Society.

http://www.siam.org/journals/sima/28-2/29083.html

[†] Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, Scotland, UK (xuerong@stams.strath.ac.uk).

or, equivalently,

$$\limsup_{t\to\infty} \frac{1}{t} \log E|x(t;\xi)|^2 \le -\gamma$$

for all $\xi \in L^2_{\mathcal{F}_0}([-\tau,0];R^n)$. *The trivial solution of equation* (1.1) *is said to be almost surely exponentially stable if there is a positive constant* $\bar{\gamma}$ *such that*

$$\limsup_{t\to\infty} \frac{1}{t} \log |x(t;\xi)| \le -\bar{\gamma} \quad a.s.$$

for all $\xi \in L^2_{\mathcal{F}_0}([-\tau,0];R^n)$.

In this paper, we shall further our study in this area. We should emphasize that the main technique employed in this paper is the well-known Razumikhin argument (see Razumikhin [13], [14]). To explain this technique, applying Itô's formula to $e^{\lambda t}|x(t) - G(x_t)|^2$, one may see that to have exponential stability in mean square, it would require that

$$E\Big(2(\phi(0) - G(\phi))^T f(t,\phi) + \text{trace}[g^T(t,\phi)g(t,\phi)]\Big) \le -\lambda E|\phi(0) - G(\phi)|^2 \tag{1.2}$$

for all $t \ge 0$ and *all* $\phi \in L^2_{\mathcal{F}_t}([-\tau,0];R^n)$. As a result, one would be forced to impose very severe restrictions on the functions $f(t,\phi)$ and $g(t,\phi)$. However, by the Razumikhin argument, one needs to require that (1.2) hold only for those $\phi \in L^2_{\mathcal{F}_t}([-\tau,0];R^n)$ satisfying

$$E|\phi(\theta)|^2 < qE|\phi(0) - G(\phi)|^2, \quad -\tau \le \theta \le 0,$$

but not necessarily for all ϕ, where $q > 1$ is a constant. Hence the restrictions on the functions $f(t,\phi)$ and $g(t,\phi)$ can be weakened considerably. This is the basic idea exploited in this paper.

This main technique of this paper is completely different from those used in our previous paper [11]. The results obtained in [11] can be applied only to a certain class of neutral stochastic functional differential equations excluding neutral stochastic differential delay equations, but the results obtained in this paper are much more general, and they especially can be used to deal with neutral stochastic differential delay equations. Moreover, in [11], we only studied the exponential stability in mean square, but in this paper, we shall also study the almost sure exponential stability. It should be pointed out that although the results established in this paper are applicable to more general neutral-type equations, for a particular class of equations discussed in [11], the results there are sharper. (Please see section 5 below for details.) Of course, this is not surprising because the results obtained by applying a particular technique to a particular equation are generally sharper than those obtained by using a general technique which is applicable to more general equations.

In this paper, the theory of existence and uniqueness of the solutions will first be introduced very briefly in section 2. The main results of this paper will be shown in sections 3 and 4, where several useful criteria will be established on the exponential stability in mean square as well as the almost sure exponential stability for the trivial solution of equation (1.1). In section 5, we shall compare our new results with the previous ones obtained in [11]. To show the power of the Razumikhin argument, the general results established in sections 3 and 4 will be applied to deal with the exponential stability of neutral stochastic differential delay equations in section 6 and of linear neutral stochastic functional differential equations in section 7.

2. Neutral stochastic functional differential equations. Throughout the paper, unless otherwise specified, we let $\tau > 0$ and $C([-\tau,0];R^n)$ denote the family of continuous functions φ from $[-\tau,0]$ to R^n with the norm $||\varphi|| = \sup_{-\tau\le\theta\le 0}|\varphi(\theta)|$, where $|\cdot|$ is the Euclidean norm in R^n. If A is a vector or matrix, its transpose is denoted by A^T. If A is a matrix, its norm $||A||$ is defined by $||A|| = \sup\{|Ax| : |x| = 1\}$ (without any confusion with $||\varphi||$). Moreover, let $w(t) = (w_1(t),\dots,w_m(t))^T$ be an m-dimensional Brownian motion defined on a complete probability space $(\Omega,\mathcal{F},P)$ with a natural filtration $\{\mathcal{F}_t\}_{t\ge 0}$ (i.e., $\mathcal{F}_t = \sigma\{w(s) : 0 \le s \le t\}$). For each $t \ge 0$, denote by $L^2_{\mathcal{F}_t}([-\tau,0];R^n)$ the family of all $\mathcal{F}_t$-measurable $C([-\tau,0];R^n)$-valued random variables $\phi = \{\phi(\theta) : -\tau \le \theta \le 0\}$ such that $\sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2 < \infty$. Also, define $L^2_{\mathcal{F}_\infty}([-\tau,0];R^n) = \bigcup_{t\ge 0} L^2_{\mathcal{F}_t}([-\tau,0];R^n)$. Obviously, $C([-\tau,0];R^n) \subset L^2_{\mathcal{F}_\infty}([-\tau,0];R^n)$.

Consider the n-dimensional neutral stochastic functional differential equation

$$d[x(t) - G(x_t)] = f(t,x_t)dt + g(t,x_t)dw(t) \tag{2.1}$$

on $t \ge 0$ with initial data $x_0 = \xi$. Here

$$G : C([-\tau,0];R^n) \to R^n, \qquad f : R_+ \times C([-\tau,0];R^n) \to R^n,$$
$$g : R_+ \times C([-\tau,0];R^n) \to R^{n\times m}$$

are all continuous functionals. Moreover, $x_t = \{x(t+\theta) : -\tau \le \theta \le 0\}$, which is regarded as a $C([-\tau,0];R^n)$-valued stochastic process, and $\xi = \{\xi(\theta) : -\tau \le \theta \le 0\} \in L^2_{\mathcal{F}_0}([-\tau,0];R^n)$. An $\mathcal{F}_t$-adapted process $x(t), -\tau \le t < \infty$ (let $\mathcal{F}_t = \mathcal{F}_0$ for $-\tau \le t \le 0$), is said to be a solution of equation (2.1) if it satisfies the initial condition and, moreover, for every $t \ge 0$,

$$x(t) - G(x_t) = \xi(0) - G(\xi) + \int_0^t f(s,x_s)ds + \int_0^t g(s,x_s)dw(s). \tag{2.1$'$}$$

To ensure the existence and uniqueness of the solution, one of the key hypotheses is the following:

(H) There is a constant $\kappa \in (0,1)$ such that

$$E|G(\phi_1) - G(\phi_2)|^2 \le \kappa \sup_{-\tau\le\theta\le 0} E|\phi_1(\theta) - \phi_2(\theta)|^2$$

for all $\phi_1,\phi_2 \in L^2_{\mathcal{F}_\infty}([-\tau,0];R^n)$.

In addition, we need further hypotheses on f and g. For example, f and g are uniformly Lipschitz continuous, or they are locally Lipschitz continuous and satisfy the linear-growth condition. Under these hypotheses, Kolmanovskii and Nosov [8] showed that there is a unique continuous solution to equation (2.1), and any moment, especially the second moment, of the solution is finite. Since the existence and uniqueness of the solution are not the main topic of this paper, we shall not discuss them in detail. All we need to do in this paper is assume that a unique solution exists and is continuous and that its second moment is finite. The solution will be denoted by $x(t;\xi)$.

3. Exponential stability in mean square. In this section, we will investigate the exponential stability in mean square for the solution of equation (2.1). For the general theory on stochastic stability, we refer the reader to Arnold [1], Friedman [2], Has'minskii [5], Mao [9, 10], or Mohammed [12]. For the stability purpose of this paper, we always assume that $G(0) = 0$, $f(t, 0) \equiv 0$, and $g(t, 0) \equiv 0$. Therefore, equation (2.1) admits a trivial solution $x(t; 0) \equiv 0$. The following Razumikhin-type theorem gives a sufficient condition for the exponential stability in mean square of this trivial solution.

THEOREM 3.1. *Assume that there is a constant $\kappa \in (0, 1)$ such that*

$$E|G(\phi)|^2 \le \kappa \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2, \quad \phi \in L^2_{\mathcal{F}_\infty}([-\tau, 0]; R^n). \tag{3.1}$$

Let $q > (1 - \sqrt{\kappa})^{-2}$. Assume furthermore that there is a $\lambda > 0$ such that

$$E\Big(2(\phi(0) - G(\phi))^T f(t, \phi) + trace[g^T(t, \phi)g(t, \phi)]\Big) \le -\lambda E|\phi(0) - G(\phi)|^2 \tag{3.2}$$

for all $t \ge 0$ and those $\phi \in L^2_{\mathcal{F}_t}([-\tau, 0]; R^n)$ satisfying

$$E|\phi(\theta)|^2 < qE|\phi(0) - G(\phi)|^2, \quad -\tau \le \theta \le 0.$$

Then for all $\xi \in L^2_{\mathcal{F}_0}([-\tau, 0]; R^n)$,

$$E|x(t; \xi)|^2 \le q(1 + \sqrt{\kappa})^2 e^{-\bar{\gamma}t} \sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2, \quad t \ge 0, \tag{3.3}$$

where

$$\bar{\gamma} = \min\left\{\lambda,\ \frac{1}{\tau}\log\left[\frac{q}{(1 + \sqrt{q\kappa})^2}\right]\right\} > 0. \tag{3.4}$$

In other words, the trivial solution of equation (2.1) *is exponentially stable in mean square.*

In order to prove this theorem, let us present two useful lemmas.

LEMMA 3.2. *Let* (3.1) *hold for some $\kappa \in (0, 1)$. Then*

$$E|\phi(0) - G(\phi)|^2 \le (1 + \sqrt{k})^2 \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2$$

for all $\phi \in L^2_{\mathcal{F}_\infty}([-\tau, 0]; R^n)$.

Proof. For any $\varepsilon > 0$,

$$\begin{aligned} E|\phi(0) - G(\phi)|^2 &\le E|\phi(0)|^2 + 2E\big(|\phi(0)||G(\phi)|\big) + E|G(\phi)|^2 \\ &\le (1 + \varepsilon)E|\phi(0)|^2 + (1 + \varepsilon^{-1})E|G(\phi)|^2 \\ &\le \big[1 + \varepsilon + \kappa(1 + \varepsilon^{-1})\big] \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2. \end{aligned}$$

Therefore, the desired result follows by taking $\varepsilon = \sqrt{\kappa}$. The proof is complete. □

LEMMA 3.3. *Let* (3.1) *hold for some $\kappa \in (0, 1)$. Let $\rho \ge 0$ and $0 < \gamma < \tau^{-1}\log(1/\kappa)$. Let $x(t)$ be a solution of equation* (2.1). *If*

$$e^{\gamma t}E|x(t) - G(x_t)|^2 \le (1 + \sqrt{\kappa})^2 \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2 \tag{3.5}$$

for all $0 \le t \le \rho$, then

$$e^{\gamma t}E|x(t)|^2 \le \frac{(1 + \sqrt{\kappa})^2}{(1 - \sqrt{\kappa e^{\gamma\tau}})^2} \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2. \tag{3.6}$$

Proof. Let $\kappa e^{\gamma\tau} < \varepsilon < 1$. For $0 \le t \le \rho$, note that

$$E|x(t) - G(x_t)|^2 \ge E|x(t)|^2 - 2E\big(|x(t)||G(x_t)|\big) + E|G(x_t)|^2$$
$$\ge (1-\varepsilon)E|x(t)|^2 - (\varepsilon^{-1} - 1)E|G(x_t)|^2.$$

Hence

$$E|x(t)|^2 \le \frac{1}{1-\varepsilon} E|x(t) - G(x_t)|^2 + \frac{\kappa}{\varepsilon} \sup_{-\tau\le\theta\le 0} E|x(t+\theta)|^2.$$

By condition (3.5), we then derive that for all $0 \le t \le \rho$,

$$e^{\gamma t}E|x(t)|^2 \le \frac{1}{1-\varepsilon} \sup_{0\le t\le\rho}\Big[e^{\gamma t}E|x(t) - G(x_t)|^2\Big] + \frac{\kappa}{\varepsilon} \sup_{0\le t\le\rho}\Big[e^{\gamma t} \sup_{-\tau\le\theta\le 0} E|x(t+\theta)|^2\Big]$$
$$\le \frac{(1+\sqrt{\kappa})^2}{1-\varepsilon} \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2 + \frac{\kappa e^{\gamma\tau}}{\varepsilon} \sup_{-\tau\le t\le\rho}\Big[e^{\gamma t}E|x(t)|^2\Big].$$

However, this holds for all $-\tau \le t \le 0$ as well. Therefore,

$$\sup_{-\tau\le t\le\rho}\Big[e^{\gamma t}E|x(t)|^2\Big] \le \frac{(1+\sqrt{\kappa})^2}{1-\varepsilon} \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2 + \frac{\kappa e^{\gamma\tau}}{\varepsilon} \sup_{-\tau\le t\le\rho}\Big[e^{\gamma t}E|x(t)|^2\Big].$$

Since $1 > \kappa e^{\gamma\tau}/\varepsilon$, we see that

$$\sup_{-\tau\le t\le\rho}\Big[e^{\gamma t}E|x(t)|^2\Big] \le \frac{\varepsilon(1+\sqrt{\kappa})^2}{(1-\varepsilon)(\varepsilon - \kappa e^{\gamma\tau})} \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2.$$

The required assertion (3.6) follows by taking $\varepsilon = \sqrt{\kappa e^{\gamma\tau}}$. The proof is complete. □

We can now begin to prove Theorem 3.1.

Proof of Theorem 3.1. First, note that $q/(1+\sqrt{q\kappa})^2 > 1$ since $q > (1-\sqrt{k})^{-2}$ and hence $\bar{\gamma} > 0$. Now fix any $\xi \in L^2_{\mathcal{F}_0}([-\tau, 0]; R^n)$ and simply write $x(t;\xi) = x(t)$. Without any loss of generality, we may assume that $\sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2 > 0$. Let $\gamma \in (0, \bar{\gamma})$ arbitrarily. It is easy to show that

$$0 < \gamma < \min\left\{\lambda,\ \frac{1}{\tau}\log\Big(\frac{1}{\kappa}\Big)\right\} \quad \text{and} \quad q > \frac{e^{\gamma\tau}}{(1-\sqrt{\kappa e^{\gamma\tau}})^2}. \tag{3.7}$$

We now claim that

$$e^{\gamma t}E|x(t) - G(x_t)|^2 \le (1+\sqrt{\kappa})^2 \sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2 \quad \text{for all } t \ge 0. \tag{3.8}$$

If so, an application of Lemma 3.3 to (3.8) yields that

$$e^{\gamma t}E|x(t)|^2 \le \frac{(1+\sqrt{\kappa})^2}{(1-\sqrt{\kappa e^{\gamma\tau}})^2} \sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2 \le q(1+\sqrt{\kappa})^2 \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2$$

for all $t \ge 0$, where we have used (3.7), and the desired result (3.3) follows by letting $\gamma \to \bar{\gamma}$. The remainder of the proof is to show (3.8) by contradiction. Suppose (3.8) is not true. Then in view of Lemma 3.2, there is a $\rho \ge 0$ such that

$$e^{\gamma t}E|x(t) - G(x_t)|^2 \le e^{\gamma\rho}E|x(\rho) - G(x_\rho)|^2 = (1+\sqrt{\kappa})^2 \sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2 \tag{3.9}$$

for all $0 \le t \le \rho$ and, moreover, there is a sequence of $\{t_k\}_{k\ge 1}$ such that $t_k \downarrow \rho$ and

$$e^{\gamma t_k} E|x(t_k) - G(x_{t_k})|^2 > e^{\gamma\rho} E|x(\rho) - G(x_\rho)|^2. \tag{3.10}$$

Applying Lemma 3.3, we derive from (3.9) that

$$\begin{aligned} e^{\gamma t} E|x(t)|^2 &\le \frac{(1+\sqrt{\kappa})^2}{(1-\sqrt{\kappa e^{\gamma\tau}})^2} \sup_{-\tau\le\theta\le 0} E|x(\theta)|^2 \\ &= \frac{e^{\gamma\rho}}{(1-\sqrt{\kappa e^{\gamma\tau}})^2} E|x(\rho) - G(x_\rho)|^2 \end{aligned}$$

for all $-\tau \le t \le \rho$. Particularly,

$$E|x(\rho+\theta)|^2 \le \frac{e^{\gamma\tau}}{(1-\sqrt{\kappa e^{\gamma\tau}})^2} E|x(\rho) - G(x_\rho)|^2 < qE|x(\rho) - G(x_\rho)|^2 \tag{3.11}$$

for all $-\tau \le \theta \le 0$, where (3.7) has been used once again. By assumption (3.2), we then have

$$E\Big(2(x(\rho) - G(x_\rho))^T f(\rho, x_\rho) + \text{trace}[g^T(\rho, x_\rho) g(\rho, x_\rho)]\Big) \le -\lambda E|x(\rho) - G(x_\rho)|^2.$$

Recalling $\gamma < \lambda$, we see by the continuity of the solution and the functionals G, f, and g (this is the standing hypothesis in this paper) that for all sufficiently small $h > 0$,

$$E\Big(2(x(t) - G(x_t))^T f(t, x_t) + \text{trace}[g^T(t, x_t) g(t, x_t)]\Big) \le -\gamma E|x(t) - G(x_t)|^2$$

if $\rho \le t \le \rho + h$. Now by Itô's formula, for all sufficiently small $h > 0$,

$$\begin{aligned} &e^{\gamma(\rho+h)}\ E|x(\rho+h) - G(x_{\rho+h})|^2 - e^{\gamma\rho} E|x(\rho) - G(x_\rho)|^2 \\ &\quad = \int_\rho^{\rho+h} e^{\gamma t}\Big[\gamma E|x(t) - G(x_t)|^2 \\ &\qquad + E\Big(2(x(t) - G(x_t))^T f(t, x_t) + \text{trace}[g^T(t, x_t) g(t, x_t)]\Big)\Big] dt \\ &\quad \le 0; \end{aligned} \tag{3.12}$$

however, this contradicts (3.10), so (3.8) must hold. The proof is now complete. □

4. Almost sure exponential stability. In this section, we discuss the almost sure exponential stability for the neutral stochastic functional differential equations. It will be shown that under the linear-growth condition, the exponential stability in mean square implies the almost sure exponential stability.

THEOREM 4.1. *Let (3.1) hold for some $\kappa \in (0,1)$. Assume that there exists a positive constant $K > 0$ such that*

$$E\big(|f(t,\phi)|^2 + \text{trace}\big[g^T(t,\phi) g(t,\phi)\big]\big) \le K \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2 \tag{4.1}$$

for all $t \ge 0$ and $\phi \in L^2_{\mathcal{F}_\infty}([-\tau, 0]; R^n)$. Assume also that the trivial solution of equation (2.1) is exponentially stable in mean square, that is, there exists a pair of positive constants γ and M such that

$$E|x(t;\xi)|^2 \le M e^{-\gamma t} \sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2, \quad t \ge 0, \tag{4.2}$$

for all $\xi \in L^2_{\mathcal{F}_0}([-\tau, 0]; R^n)$. *Then*

$$\limsup_{t\to\infty} \frac{1}{t} \log |x(t;\xi)| \le -\frac{\bar\gamma}{2} \quad a.s., \tag{4.3}$$

where $\bar\gamma = \min\{\gamma, \tau^{-1}\log(1/\kappa)\}$, *that is, the trivial solution of equation* (2.1) *is also almost surely exponentially stable. In particular, if* (3.1), (3.2), *and* (4.1) *hold, then the trivial solution of equation* (2.1) *is almost surely exponentially stable.*

To prove the theorem, we need to present another lemma which is very useful in the study of the almost sure exponential stability of neutral stochastic functional differential equations.

Lemma 4.2. *Assume that there exists a constant* $\kappa \in (0, 1)$ *such that*

$$|G(\varphi)|^2 \le \kappa \sup_{-\tau\le\theta\le 0} |\varphi(\theta)|^2, \quad \varphi \in C([-\tau, 0]; R^n). \tag{4.4}$$

Let $z : [-\tau, \infty) \to R^n$ *be a continuous function. Let* $0 < \gamma < \tau^{-1}\log(1/\kappa)$ *and* $H > 0$. *If*

$$|z(t) - G(z_t)|^2 \le He^{-\gamma t} \quad \textit{for all } t \ge 0, \tag{4.5}$$

then

$$\limsup_{t\to\infty} \frac{1}{t} \log |z(t)| \le -\frac{\gamma}{2}. \tag{4.6}$$

Proof. Choose any $\varepsilon \in (\kappa e^{\gamma\tau}, 1)$. In the same way as in the proof of Lemma 3.3, we can show that for any $T > 0$,

$$\sup_{0\le t\le T} \Big[e^{\gamma t}|z(t)|^2\Big] \le \frac{H}{1-\varepsilon} + \frac{\kappa e^{\gamma\tau}}{\varepsilon} \sup_{-\tau\le t\le T} \Big[e^{\gamma t}|z(t)|^2\Big].$$

It then follows that

$$\Big(1 - \frac{\kappa e^{\gamma\tau}}{\varepsilon}\Big) \sup_{0\le t\le T} \Big[e^{\gamma t}|z(t)|^2\Big] \le \frac{H}{1-\varepsilon} + \frac{\kappa e^{\gamma\tau}}{\varepsilon} \sup_{-\tau\le t\le 0} |z(t)|^2.$$

Consequently,

$$\limsup_{t\to\infty} \frac{1}{t} \log |z(t)| \le -\frac{\gamma}{2},$$

as required. The proof is complete. □

Proof of Theorem 4.1. First, note that condition (4.1) implies condition (4.4) since $C[-\tau; 0]; R^n) \subset L^2_{\mathcal{F}_\infty}([-\tau, 0]; R^n)$. Now fix any initial data ξ and write the solution $x(t;\xi) = x(t)$ simply. By the well-known Doob martingale inequality (cf. Karatzas and Shreve [6]), the Hölder inequality, and condition (4.2), we can easily derive that for any integer $k \ge 1$,

(4.7)

$$\begin{aligned} &E\Big(\sup_{0\le\theta\le\tau} |x(k\tau+\theta) - G(x_{k\tau+\theta})|^2\Big) \\ &\le 3E|x(k\tau) - G(x_{k\tau})|^2 + 3K(\tau+4)\int_{k\tau}^{(k+1)\tau} \Big(\sup_{-\tau\le\theta\le 0} E|x(s+\theta)|^2\Big)ds \\ &\le \Big(6M(1+\kappa)e^{-\bar\gamma(k\tau-\tau)} + 3K(\tau+4)M\int_{k\tau}^{(k+1)\tau} e^{-\bar\gamma(s-\tau)}ds\Big)\Big[\sup_{-\tau\le\theta\le 0} E|\xi(\theta)|^2\Big] \\ &\le Ce^{-\bar\gamma k\tau}, \end{aligned}$$

where $C = 3Me^{\bar{\gamma}\tau}\big[2(1+\kappa)+K(\tau+4)\big]\sup_{-\tau\le\theta\le0} E|\xi(\theta)|^2$. Let $\varepsilon \in (0, \bar{\gamma})$ be arbitrary. It then follows from (4.7) that

$$P\left(\omega : \sup_{0\le\theta\le\tau} |x(k\tau+\theta) - G(x_{k\tau+\theta})|^2 > e^{-(\bar{\gamma}-\varepsilon)k\tau}\right) \le Ce^{-\varepsilon k\tau}.$$

In view of the well-known Borel–Cantelli lemma, we see that for almost all $\omega \in \Omega$,

$$\sup_{0\le\theta\le\tau} |x(k\tau+\theta) - G(x_{k\tau+\theta})|^2 \le e^{-(\bar{\gamma}-\varepsilon)k\tau} \tag{4.8}$$

holds for all but finitely many k. Hence for all $\omega \in \Omega$ excluding a P-null set, there exists a $k_o(\omega)$ for which (4.8) holds whenever $k \ge k_o$. In other words, for almost all $\omega \in \Omega$,

$$|x(t) - G(x_t)|^2 \le e^{-(\bar{\gamma}-\varepsilon)(t-\tau)} \quad \text{if } t \ge k_o\tau.$$

However, $|x(t) - G(x_t)|^2$ is finite on $[0,\ k_o\tau]$. Therefore, for almost all $\omega \in \Omega$, there exists a finite number $H = H(\omega)$ such that

$$|x(t) - G(x_t)|^2 \le He^{-(\bar{\gamma}-\varepsilon)t} \quad \text{for all } t \ge 0.$$

An application of Lemma 4.2 now yields

$$\limsup_{t\to\infty} \frac{1}{t}\log|x(t)| \le -\frac{\bar{\gamma}-\varepsilon}{2} \quad \text{a.s.},$$

and the desired result (4.3) follows by letting $\varepsilon \to 0$. The proof is complete. □

5. Comparison with existing results.

Recently, in [11], we studied the exponential stability in mean square for a class of neutral stochastic functional differential equations using a completely different technique from the one in this paper. The aim of this section is to compare our previous results in [11] with our new results in this paper. The equation studied in [11] is of the form

$$d[x(t) - G(x_t)] = [f_1(t, x(t)) + f_2(t, x_t)]dt + g(t, x_t)dw(t) \tag{5.1}$$

on $t \ge 0$ with initial data $x_0 = \xi$, where $f_1 : R_+ \times R^n \to R^n,\ f_2 : R_+ \times C([-\tau, 0]; R^n) \to R^n$, and G and g are the same as before. Let us first state a useful result.

Theorem 5.1. *Let* (3.1) *hold. Assume that there are two positive constants* λ_1 *and* λ_2 *such that*

$$\begin{aligned} &E\Big(2(\phi(0) - G(\phi\ ))^T[f_1(t, \phi(0)) + f_2(t, \phi)] + trace[g^T(t, \phi)g(t, \phi)]\Big) \\ &\qquad \le -\lambda_1 E|\phi(0)|^2 + \lambda_2 \sup_{-\tau\le\theta\le0} E|\phi(\theta)|^2 \end{aligned} \tag{5.2}$$

for all $t \ge 0$ *and* $\phi \in L^2_{\mathcal{F}_\infty}([-\tau, 0]; R^n)$. *If*

$$0 < \kappa < \frac{1}{4} \quad \textit{and} \quad \lambda_1 > \frac{\lambda_2}{(1 - 2\sqrt{\kappa})^2}, \tag{5.3}$$

then the trivial solution of equation (5.1) *is exponentially stable in mean square.*

Proof. By condition (5.3), we can choose q such that

$$\frac{1}{\kappa} > q > \frac{1}{(1-\sqrt{\kappa})^2} \quad \text{and} \quad \lambda_1 > \frac{\lambda_2 q}{(1-\sqrt{\kappa q})^2}. \tag{5.4}$$

By defining $f(t,\varphi) = f_1(t,\varphi(0)) + f_2(t,\varphi)$ for $t \geq 0$ and $\varphi \in C([-\tau,0];R^n)$, equation (5.1) can be written as equation (2.1), so all that we need to do is verify condition (3.2). To do so, let $t \geq 0$ and $\phi \in L^2_{\mathcal{F}_t}([-\tau,0];R^n)$, satisfying

$$E|\phi(\theta)|^2 < qE|\phi(0) - G(\phi)|^2, \quad -\tau \leq \theta \leq 0.$$

Note that for any $\varepsilon > 0$,

$$-E|\phi(0)|^2 \leq -\frac{1}{1+\varepsilon}E|\phi(0) - G(\phi)|^2 + \frac{1}{\varepsilon}E|G(\phi)|^2. \tag{5.5}$$

It then follows from (5.2) and (5.5) that

$$\begin{aligned} &E\Big(2(\phi(0) - G(\phi))^T[f_1(t,\phi(0)) + f_2(t,\phi)] + \text{trace}[g^T(t,\phi)g(t,\phi)]\Big) \\ &\quad\leq -\lambda_1 E|\phi(0)|^2 + \lambda_2 \sup_{-\tau\leq\theta\leq 0} E|\phi(\theta)|^2 \\ &\quad\leq -\Big[\lambda_1\Big(\frac{1}{1+\varepsilon} - \frac{\kappa q}{\varepsilon}\Big) - \lambda_2 q\Big]E|\phi(0) - G(\phi)|^2. \end{aligned} \tag{5.6}$$

In particular, choose $\varepsilon = \sqrt{\kappa q}/(1-\sqrt{\kappa q})$ and hence

$$\Big[\lambda_1\Big(\frac{1}{1+\varepsilon} - \frac{\kappa q}{\varepsilon}\Big) - \lambda_2 q\Big] = \lambda_1(1-\sqrt{\kappa q})^2 - \lambda_2 q > 0,$$

where we have used (5.4). In other words, condition (3.2) is satisfied and hence the conclusion follows from Theorem 3.1. The proof is complete. □

To compare this result with one in our previous paper [11], let us introduce another new notation $\mathcal{W}([-\tau,0];R_+)$, which is the family of all Borel-measurable bounded nonnegative functions $\eta(\theta)$ defined on $-\tau \leq \theta \leq 0$ such that $\int_{-\tau}^0 \eta(\theta)d\theta = 1$. In [11], conditions (3.1) and (5.2) were strengthened as follows: There is a constant $\kappa \in (0,1)$ and a function $\eta_1 \in \mathcal{W}([-\tau,0];R_+)$ such that

$$|G(\varphi)|^2 \leq \kappa \int_{-\tau}^0 \eta_1(\theta)|\varphi(\theta)|^2 d\theta \quad \text{for all } \varphi \in C([-\tau,0];R^n); \tag{5.7}$$

moreover, there exists a function $\eta_2(.) \in \mathcal{W}([-\tau,0];R_+)$ and two positive constants λ_1 and λ_2 such that

$$\begin{aligned} &2(\varphi(0) - G(\varphi\ ))^T[f_1(t,\varphi(0)) + f_2(t,\varphi)] + \text{trace}[g^T(t,\varphi)g(t,\varphi)] \\ &\quad\leq -\lambda_1|\varphi(0)|^2 + \lambda_2\int_{-\tau}^0 \eta_2(\theta)|\varphi(\theta)|^2 d\theta \end{aligned} \tag{5.8}$$

for all $t \geq 0$ and $\varphi \in C([-\tau,0];R^n)$. These two conditions are indeed stronger than (3.1) and (5.2), respectively. For example, if (5.7) holds, then for any $\phi \in$

$L^2_{\mathcal{F}_\infty}([-\tau,0];R^n)$,

$$\begin{aligned} E|G(\phi)|^2 &\le \kappa \int_{-\tau}^0 \eta_1(\theta) E|\phi(\theta)|^2 d\theta \\ &\le \kappa \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2 \int_{-\tau}^0 \eta_1(\theta) d\theta = \kappa \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2, \end{aligned}$$

that is, (3.1) holds. However, what we gained in [11] with this price paid is the following sharper result.

Theorem 5.2 (Mao [11]). *Let* (5.7) *and* (5.8) *hold. If* $\lambda_1 > \lambda_2$ *and* $\kappa \in (0,1)$, *then the trivial solution of equation* (5.1) *is exponentially stable in mean square.*

Obviously, $\lambda_1 > \lambda_2$ and $\kappa \in (0,1)$ are much sharper than (5.3). The disadvantage of Theorem 5.2 is that (5.7) and (5.8) are somehow too restricted. For instance, Theorem 5.2 is not applicable to the case of neutral stochastic differential delay equations. However, Theorem 5.1 can be applied to deal with the delay case easily. Let us now turn to this topic.

6. Neutral stochastic differential delay equations. As an application, let us apply the theory established in the previous sections to deal with the exponential stability of neutral stochastic differential delay equations of the form

$$d[x(t) - \bar{G}(x(t-\tau))] = \bar{f}(t, x(t), x(t-\tau))dt + \bar{g}(t, x(t), x(t-\tau))dw(t) \tag{6.1}$$

on $t \ge 0$ with initial data $x_0 = \xi$, where $\bar{G} : R^n \to R^n$, $\bar{f} : R_+ \times R^n \times R^n \to R^n$, and $\bar{g} : R_+ \times R^n \times R^n \to R^{n\times m}$. As before, assume that equation (6.1) has a unique global solution denoted by $x(t;\xi)$ and, moreover, $\bar{G}(0) = 0$, $\bar{f}(t,0,0) = 0$, and $\bar{g}(t,0,0) = 0$. We now employ Theorem 5.1 to establish one useful corollary.

Corollary 6.1. *Assume that there is a constant* $\kappa \in (0,1)$ *such that*

$$|\bar{G}(x)|^2 \le \kappa |x|^2, \quad x \in R^n. \tag{6.2}$$

Assume also that there are two positive constants λ_1 *and* λ_2 *such that*

$$2(x - \bar{G}(y))^T \bar{f}(t,x,y) + \mathrm{trace}[\bar{g}^T(t,x,y)\bar{g}(t,x,y)] \le -\lambda_1 |x|^2 + \lambda_2 |y|^2 \tag{6.3}$$

for all $t \ge 0$ *and* $x, y \in R^n$. *If*

$$0 < \kappa < \frac{1}{4} \quad \textit{and} \quad \lambda_1 > \frac{\lambda_2}{(1-2\sqrt{\kappa})^2}, \tag{6.4}$$

then the trivial solution of equation (6.1) *is exponentially stable in mean square.*

This corollary follows directly from Theorem 5.1 since equation (6.1) can be written as equation (5.1) by defining

$$G(\varphi) = \bar{G}(\varphi(-\tau)), \qquad f_1(t,x) = \bar{f}(t,x,0),$$

$$f_2(t,\varphi) = -\bar{f}(t,\varphi(0),0) + \bar{f}(t,\varphi(0),\varphi(-\tau)), \qquad g(t,\varphi) = \bar{g}(t,\varphi(0),\varphi(-\tau))$$

for $t \ge 0$, $x \in R^n$ and $\varphi \in C([-\tau,0];R^n)$. Of course, we can directly apply Theorems 3.1 and 4.1 to obtain a more general result. For this purpose, let us introduce another new notation $L^2_{\mathcal{F}}(\Omega;R^n)$, which is the family of R^n-valued $\mathcal{F}$-measurable random variables X such that $E|X|^2 < \infty$.

THEOREM 6.2. *Let* (6.2) *hold with* $\kappa \in (0,1)$. *Let* $q > (1-\sqrt{\kappa})^{-2}$. *Assume that there is a constant* $\lambda > 0$ *such that*

$$E\Big(2(X-\bar{G}(Y))^T\ \bar{f}(t,X,Y) + \text{trace}[\bar{g}^T(t,X,Y)\bar{g}(t,X,Y)]\Big)$$

$$\le -\lambda E|X-\bar{G}(Y)|^2 \tag{6.5}$$

for all $t \ge 0$ *and those* $X, Y \in L^2_{\mathcal{F}}(\Omega; R^n)$ *satisfying* $E|Y|^2 < qE|X-\bar{G}(Y)|^2$. *Then the trivial solution of equation* (6.1) *is exponentially stable in mean square. Furthermore, if there is a positive constant* K *such that*

$$|\bar{f}(t,x,y)|^2 + \text{trace}[\bar{g}^T(t,x,y)\bar{g}(t,x,y)] \le K(|x|^2+|y|^2), \quad x,y \in R^n, \tag{6.6}$$

then the trivial solution of equation (6.1) *is also almost surely exponentially stable.*

This theorem follows directly from Theorems 3.1 and 4.1 since equation (6.1) can be written as equation (2.1) by defining

$$G(\varphi) = \bar{G}(\varphi(-\tau)), \qquad f(t,\varphi) = \bar{f}(t,\varphi(0),\varphi(-\tau)), \qquad g(t,\varphi) = \bar{g}(t,\varphi(0),\varphi(-\tau))$$

for $t \ge 0$ and $\varphi \in C([-\tau,0]; R^n)$.

7. Linear neutral stochastic functional differential equations. As another application, let us consider a linear neutral stochastic functional differential equation

$$d[x(t) - G(x_t)] = [-Ax(t) + B_0(x_t)]dt + \sum_{i=1}^m B_i(x_t)dw_i(t) \tag{7.1}$$

on $t \ge 0$ with initial data $x_0 = \xi$. Here A is an $n \times n$ constant matrix and

$$G(\varphi) = \int_{-\tau}^0 d\gamma(\theta)\varphi(\theta), \qquad B_i(\varphi) = \int_{-\tau}^0 d\beta_i(\theta)\varphi(\theta)$$

for $\varphi \in C([-\tau,0]; R^n)$, $0 \le i \le m$, where $\gamma(\theta) = (\gamma^{kl}(\theta))_{n\times n}$, $\beta_i(\theta) = (\beta_i^{kl}(\theta))_{n\times n}$ and all $\gamma^{kl}(\theta)$ and $\beta_i^{kl}(\theta)$ are functions of bounded variation on $-\tau \le \theta \le 0$. Let $V_{\gamma^{kl}}(\theta)$ denote the total variations of γ^{kl} on the interval $[-\tau,\theta]$ and let $V_\gamma(\theta) = ||V_{\gamma^{kl}}(\theta)||$. We can define $V_{\beta_i}(\theta)$ similarly. In particular, let

$$\hat{\gamma} = V_\gamma(0) \quad \text{and} \quad \hat{\beta}_i = V_{\beta_i}(0), \quad 0 \le i \le m.$$

Let us now impose the first assumption:

$$0 < \hat{\gamma} < \frac{1}{2}. \tag{7.2}$$

Then for any $\phi \in L^2_{\mathcal{F}_\infty}([-\tau,0]; R^n)$,

$$E|G(\varphi)|^2 \le \hat{\gamma}E\int_{-\tau}^0 dV_\gamma(\theta)|\varphi(\theta)|^2 \le \hat{\gamma}^2 \sup_{-\tau\le\theta\le0} E|\varphi(\theta)|^2. \tag{7.3}$$

In other words, (3.1) is satisfied with $\kappa = \hat{\gamma}^2$. Moreover,

$$2E\big[|\phi(0)||G(\phi)|\big] \le \frac{\hat{\gamma}}{1-2\hat{\gamma}}E|\phi(0)|^2 + \frac{1-2\hat{\gamma}}{\hat{\gamma}}E|G(\phi)|^2$$

$$\le \frac{\hat{\gamma}}{1-2\hat{\gamma}}E|\phi(0)|^2 + \hat{\gamma}(1-2\hat{\gamma}) \sup_{-\tau\le\theta\le0} E|\varphi(\theta)|^2. \tag{7.4}$$

Similarly, one can show that

$$2E\big[|\phi(0)||B_0(\phi)|\big] \le \frac{\hat{\beta}_0}{1-2\hat{\gamma}} E|\phi(0)|^2 + \hat{\beta}_0(1-2\hat{\gamma}) \sup_{-\tau\le\theta\le 0} E|\varphi(\theta)|^2, \tag{7.5}$$

$$2E\big[|G(\phi)||B_0(\phi)|\big] \le 2\hat{\gamma}\hat{\beta}_0 \sup_{-\tau\le\theta\le 0} E|\varphi(\theta)|^2, \tag{7.6}$$

and

$$\sum_{i=1}^{m} E|B_i(\phi)|^2 \le \left[\sum_{i=1}^{m} \hat{\beta}_i^2\right] \sup_{-\tau\le\theta\le 0} E|\varphi(\theta)|^2. \tag{7.7}$$

Let $\lambda_{\min}(A+A^T)$ denote the smallest eigenvalue of $A+A^T$. Using (7.4)–(7.7), we then see that

$$\begin{aligned} E&\left(2(\phi(0)-G(\phi))^T[-A\phi(0)+B_0(\phi)] + \sum_{i=1}^{m} E|B_i(\phi)|^2\right) \\ &\le -\left[\lambda_{\min}(A+A^T) - \frac{\hat{\gamma}||A||+\hat{\beta}_0}{1-2\hat{\gamma}}\right] E|\phi(0)|^2 \\ &\quad + \left[(\hat{\gamma}||A||+\hat{\beta}_0)(1-2\hat{\gamma}) + 2\hat{\gamma}\hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i^2\right] \sup_{-\tau\le\theta\le 0} E|\phi(\theta)|^2. \end{aligned} \tag{7.8}$$

To close this paper, we apply Theorems 5.1 and 4.1 and conclude the following corollary.

Corollary 7.1. *Let* (7.2) *hold. If*

$$\lambda_{\min}(A+A^T) > \frac{2(\hat{\gamma}||A||+\hat{\beta}_0)}{1-2\hat{\gamma}} + \frac{1}{(1-2\hat{\gamma})^2}\left[2\hat{\gamma}\hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i^2\right],$$

then the trivial solution of equation (7.1) *is exponentially stable in mean square and is also almost surely exponentially stable.*

Acknowledgment. The author would like to thank the anonymous referee for helpful suggestions.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1972.

[2] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 2, Academic Press, New York, 1976.

[3] J. R. Haddock, T. Krisztin, J. Terjéki, and J. H. Wu, *An invariance principle of Lyapunov–Razumikhin type for neutral functional differential equations*, J. Differential Equations, 107 (1994), pp. 395–417.

[4] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, Berlin, 1993.

[5] R. Z. Has'minskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Groningen, the Netherlands, 1981.

[6] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, Berlin, 1991.

[7] V. B. Kolmanovskii and A. Myshkis, *Applied Theory of Functional Differential Equations*, Kluwer Academic Publishers, Norwell, MA, 1992.

[8] V. B. Kolmanovskii and V. R. Nosov, *Stability of Functional Differential Equations*, Academic Press, New York, 1986.

[9] X. Mao, *Stability of Stochastic Differential Equations with Respect to Semimartingales*, Longman Scientific and Technical, Harlow, UK, 1991.

[10] X. Mao, *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.

[11] X. Mao, *Exponential stability in mean square of neutral stochastic differential functional equations*, Systems Control Lett., 26 (1995), pp. 245–251.

[12] S. E. A. Mohammed, *Stochastic Functional Differential Equations*, Longman Scientific and Technical, Harlow, UK, 1986.

[13] B. S. Razumikhin, *On the stability of systems with a delay*, Prikl. Mat. Mekh., 20 (1956), pp. 500–512.

[14] B. S. Razumikhin, *Application of Lyapunov's method to problems in the stability of systems with a delay*, Automat. i Telemekh., 21 (1960), pp. 740–749.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 402–433, March 1997

010

AN EXTENSION OF THE STABILITY INDEX FOR TRAVELING-WAVE SOLUTIONS AND ITS APPLICATION TO BIFURCATIONS[*]

SHUNSAKU NII[†]

Abstract. We treat the stability index for traveling-wave solutions of one-dimensional reaction-diffusion equations due to Alexander, Gardner, and Jones [*J. Reine Angew. Math.*, 410 (1990), pp. 167–212]. An extension of the stability index which makes the index robust to perturbation is given and, using the extension, an additive formula for a gluing bifurcation of traveling waves is proven. We also consider certain heteroclinic bifurcations as an application, some specific examples of which are discussed.

Key words. traveling wave, stability index, heteroclinic bifurcation

AMS subject classifications. 34, 35

PII. S003614109427878X

1. Introduction.

We often encounter the following type of equations as model equations in chemistry, mathematical biology, and other areas concerning formation of spatial patterns; these kinds of equations are called reaction-diffusion equations:

$$\begin{cases} \frac{\partial u_1}{\partial t} = d_1 \Delta u_1 + f_1(u_1, \ldots, u_n), \\ \quad\vdots \\ \frac{\partial u_n}{\partial t} = d_n \Delta u_n + f_n(u_1, \ldots, u_n), \end{cases} \tag{1.1}$$

where $d_i \geq 0$ represents—as a model equation of a chemical reaction, for instance—the spatial diffusion rate of chemicals and f_i models their production.

Stable steady-state constant solutions of such systems correspond to homogeneous equilibrium states. On the other hand, pattern formations which appear in transition processes are attracting interest, and much research has been devoted to them. For systems with a one-dimensional space variable, traveling-wave solutions are especially important from this point of view. In this paper, we deal with their stability and bifurcations.

Consider a one-dimensional reaction-diffusion system

$$u_t = Bu_{xx} + F(u), \tag{1.2}$$

where $x, t \in \mathbb{R}$, $u(x,t) \in \mathbb{R}^n$, B is an $n \times n$ positive diagonal matrix, and $F\colon \mathbb{R}^n \to \mathbb{R}^n$. Set $\xi = x - \theta t$ and rewrite this equation on a moving frame (ξ, t); then we have

$$u_t = Bu_{\xi\xi} + \theta u_\xi + F(u). \tag{1.3}$$

In the x-coordinate, a steady-state solution of equation (1.3) is a solution which translates with constant velocity θ, preserving the solution's profile. This is a traveling-wave

[*] Received by the editors December 16, 1994; accepted for publication (in revised form) November 9, 1995. This work was supported by the JSPS Research Fellowships for Young Scientists and Science Foundation of the Ministry of Education of Japan grant 3078.

http://www.siam.org/journals/sima/28-2/27878.html

[†] Department of Mathematics, Faculty of Science, Saitama University, 255 Shimo-Ohkubo, Urawa-shi 338, Japan (snii@rimath.saitama-u.ac.jp).

solution of system (1.2). In this paper, we restrict our attention to the traveling waves satisfying the boundary condition

$$\lim_{x\to\pm\infty} u(x,t) = u_\pm, \tag{1.4}$$

where we assume that $u \equiv u_\pm$ are stable steady-state constant solutions of (1.2), i.e., $u(x,t) = u(\xi)$ is a solution of the ordinary differential equation.

$$Bu_{\xi\xi} + \theta u_\xi + F(u) = 0 \tag{1.5}$$

with $\lim_{\xi\to\pm\infty} u(\xi) = u_\pm$. This means that $(u(\xi), u'(\xi))$ is a heteroclinic solution connecting $(u_-, 0)$ and $(u_+, 0)$ of the first-order system

$$\begin{cases} u' = v, \\ v' = -B^{-1}F(u) - \theta B^{-1}v \end{cases} \quad ({}' = \tfrac{d}{d\xi}). \tag{1.6}$$

This kind of wave corresponds to the movement of a transition layer between two equilibrium states when u_- is different from u_+, whereas it corresponds to the propagation of a pulse when u_- equals to u_+.

The existence problem of such traveling waves or, equivalently, heteroclinic (homoclinic) orbits of (1.6) already commands a large body of literature, and it is also one of the main sources of motivation for the development of bifurcation theory for homoclinic or heteroclinic orbits of vector fields.

However, not all traveling waves correspond to observed phenomena which are described by the equations. Physical systems always suffer from noise which comes from outside the system. Therefore, waves represented by unstable solutions, which are destroyed by small perturbations of their initial conditions, cannot sustain themselves in the systems in the real world. That is why we need stability analysis; in other words, we have to know stable solutions which, even if they are perturbed slightly, recover to the original solutions as time goes on.

We usually use the technique of linear stability; that is, we linearize the equation about the wave under study and investigate whether this linearized operator on the appropriate space has no eigenvalue with a positive real part. This is because, in the context of semilinear parabolic equations, it is well known that stability for the linear problem implies the same for the nonlinear problem under an appropriate setting. Therefore, the eigenvalue problem is the main issue for the stability analysis of traveling waves. See Alexander, Gardner, and Jones [1], and for more details, see Bates and Jones [4] and Henry [9].

One way to approach the eigenvalue problem is to investigate the Evans function, which is an analytic function on the complex plane derived from the eigenvalue problem and vanishes at the eigenvalues (see Evans [6], [7], Jones [12], and Yanagida [20]). This function is the basis of the stability index due to Alexander, Gardner, and Jones [1], which is the main subject of this paper. In fact, it is an algebraically dressed refinement of the Evans function.

The aim of this paper is to determine the stability of traveling waves that are generated by certain bifurcations.

First, we present an extension of the stability index.

Consider a system depending on a parameter $\mu \in \mathbb{R}$,

$$u_t = Bu_{xx} + F(u;\mu). \tag{1.7}$$

Assume that at $\mu = 0$ this system has a traveling-wave solution $u_0(\xi)$ with velocity θ, i.e., $u_0(\xi)$ satisfies

$$Bu_{\xi\xi} + \theta u_\xi + F(u;\mu) = 0 \tag{1.8}$$

for $\xi = x - \theta t$. The stability index is defined for u_0. For $\mu \neq 0$, (1.7) generically does not have a traveling wave, so the stability index is not defined. However, equation (1.8) or the following system corresponding to it should retain some information for $\mu \approx 0$, and it is important to obtain this information when (1.8) (or (1.9)) undergoes various bifurcations:

$$\begin{cases} u' = v, \\ v' = -B^{-1}F(u;\mu) - \theta B^{-1}v \end{cases} \qquad ({}' = \tfrac{d}{d\xi}). \tag{1.9}$$

In fact, for each μ, (1.9) has a solution $(u(\xi;\mu), u'(\xi;\mu))$ that is "near" $(u_0(\xi), u_0'(\xi))$, and we can make use of this. The stability index is originally defined as the first Chern number of a certain vector bundle over the two-dimensional sphere, but we regard it as an element of the two-dimensional homotopy group of the complex projective space (Theorem 2.2). We prove that this definition has a kind of robustness; namely, we can define the index for $u(\xi;\mu)$ as an element of the relative homotopy group, and it inherits the information of the index defined for $u_0(\xi)$ (Theorem 3.1).

Next, an additive formula (Theorem 3.2) of the index of the gluing bifurcation [8] of heteroclinic orbits of (1.9) is proven as an application of the extension of the stability index.

Sometimes we see gluing bifurcations of traveling waves—two traveling waves are glued with each other at the ends of $\xi \to \pm\infty$ and a new traveling wave is produced. At first, Yanagida and Maginu [21] treated this kind of problem for pulses with oscillating tails and where the equilibrium of (1.9) with which the pulses are associated has a one-dimensional unstable manifold. In this case, the definition and calculation of the Evans function was relatively easy. Later, Alexander and Jones [2], [3] dealt with the same problem when the unstable manifold is two dimensional. In [2], an additive formula for the bifurcation is proven using cobordism. Our result provides another way of proving it.

The additive formula tells us that for a wave generated by a gluing bifurcation of two waves, the linearized operator possesses two eigenvalues near the origin; one is at the origin corresponding to the translation of the wave, so the sign of the other one determines the stability. Alexander and Jones [2] showed that the direction of the intersection of stable and unstable manifolds of equilibria of (1.9) determines the sign (Proposition 4.1). This is also a generalization of the works of Evans [7], Jones [12], Yanagida [20], and Yanagida and Maginu [21]. In those works, the geometrical meaning of direction was clear. In the general situation, however, we must define the direction properly.

Besides the additive formula, other applications of the extension based on its robustness are expected. For example, a similar index is defined in [19] to detect the existence of N-homoclinic bifurcations, and there may be some relation between these two indices.

The remainder of this paper is devoted to an application to a heteroclinic bifurcation. A generic two-parameter family of vector fields with two heteroclinic orbits in a row undergoes a gluing bifurcation that produces a heteroclinic orbit that stays in some neighborhood of the concatenation of original two heteroclinic orbits (see Kokubu [17]). This type of bifurcation is also observed in some reaction-diffusion equations. We apply the preceding argument to this kind of bifurcation. In this case, we must determine the direction of the intersections to apply to actual examples (Theorem 4.3). This is not an easy task in general situations, but we can determine the stability for the case of a heteroclinic loop. More precisely, assume that system (1.9) has two equilibria $(u_1, 0)$ and $(u_2, 0)$, and assume that for a certain parameter value, there exist two heteroclinic orbits from $(u_1, 0)$ to $(u_2, 0)$ and from $(u_2, 0)$ to $(u_1, 0)$ forming a loop called a heteroclinic loop. Then the system undergoes a gluing bifurcation, producing a homoclinic orbit with respect to $(u_1, 0)$. In this situation, we show that the stability of the traveling wave corresponding to the homoclinic orbit is determined by the bifurcation diagram (Theorem 4.4); we show the relation between bifurcation diagrams and stability in Figure 5.7. This is another main theorem of this paper.

This paper is divided into five parts. In section 2, we briefly summarize the definition of the stability index and translate it in order to treat it via our methods. In section 3, we prove that the index can be extended and give a proof of the additive formula. Sections 4 and 5 are devoted to an application to a heteroclinic bifurcation and its examples.

2. The stability index.

First, we briefly summarize the construction of stability indices for traveling-wave solutions of one-dimensional reaction-diffusion equations due to Alexander, Gardner, and Jones. For details and proofs, see [1] and the references therein.

2.1. Traveling waves and their stability.

Consider one-dimensional reaction-diffusion equations of the following form:

$$u_t = Bu_{xx} + F(u). \tag{2.1}$$

Here $u(x,t) \in \mathbb{R}^n$, $x, t \in \mathbb{R}$, B : positive diagonal $n \times n$ matrix, and $F : C^2$ with derivatives through order 2 bounded on $\mathbb{R}^n$.

We get on the moving frame (ξ, t), where $\xi = x - \theta t$ and the velocity θ is a constant. In this coordinate, (2.1) is written as follows:

$$u_t = Bu_{\xi\xi} + \theta u_\xi + F(u). \tag{2.2}$$

A traveling-wave solution of (2.1) is a stationary solution $u_0(\xi) = u_0(x - \theta t)$ of (2.2) for some θ; in other words, it is a solution of (2.1) which translates at some constant velocity θ preserving its shape.

In what follows, we assume the existence of a traveling-wave solution satisfying the following assumptions.

(T.W.) There exist $C > 0$, $a > 0$, and $u_\pm \in \mathbb{R}^n$ with $F(u_\pm) = 0$ such that

1. $|u_0(\xi) - u_+| \leq Ce^{-a\xi}$ $(\xi > 0)$,
2. $|u_0(\xi) - u_-| \leq Ce^{a\xi}$ $(\xi < 0)$, and
3. $|u_0'(\xi)| \leq Ce^{-a|\xi|}$ ($\xi \in \mathbb{R}$, where $' = \frac{d}{d\xi}$).

Next, we define stability of traveling waves.

Definition 2.1. *A traveling wave $u_0(\xi)$ is asymptotically stable relative to* (2.2) *if there exists a neighborhood N of u_0 in $BU(\mathbb{R}, \mathbb{R}^n)$ such that each solution $u(\xi, t)$ of* (2.2) *that starts in N at $t = 0$ satisfies*

$$\|u(\xi,t) - u_0(\xi + k)\|_\infty \to 0 \quad (t \to +\infty) \tag{2.3}$$

for some $k \in \mathbb{R}$ depending on $u(\xi, t)$, where $BU(\mathbb{R}, \mathbb{R}^n) = \{v : \mathbb{R} \to \mathbb{R}^n |$ bounded uniformly continuous$\}$.

Remark 2.1. If $u_0(\xi)$ is a traveling wave, then so is $u_0(\xi + k)$.

To find the stability of $u_0(\xi)$, we linearize (2.2) about $u_0(\xi)$,

$$P_t = LP := BP_{\xi\xi} + \theta P_\xi + DF(u_0(\xi))P, \tag{2.4}$$

and consider the eigenvalue problem for L:

$$LP = \lambda P. \tag{2.5}$$

In order to determine the stability of u_0, it suffices to study the eigenvalue problem (2.5) because of the following well-known fact.

Fact (see [1], [4], and [9]). *$u_0(\xi)$ is asymptotically stable if the spectrum $\sigma(L)$ of L satisfies the following:*

1. *there exists $\beta < 0$ such that $\sigma(L) \setminus \{0\} \subset \{\lambda | \mathrm{Re}\lambda < \beta\}$;*
2. *0 is a simple eigenvalue* (0 *is an eigenvalue corresponding to translation*)*.*

Here we have the following for $\sigma(L)$.

Proposition 2.1. *If $u_\pm$ are stable for* (2.1)*, then there exists a simple closed curve K and a constant $\beta < 0$ such that*

$$\sigma(L) \cap \{\lambda | \mathrm{Re}\lambda > \beta\} \subset K^\circ, \tag{2.6}$$

where K° is the interior enclosed by K. Moreover, $\sigma(L) \cap K^\circ$ consists of isolated eigenvalues with finite multiplicity.

With this proposition, we conclude that $u_0(\xi)$ is stable if L has no eigenvalue in K° other than the simple one at 0.

Henceforth, we construct an index to detect eigenvalues of L inside K°.

2.2. Construction of the stability index. The eigenvalue problem

$$(L - \lambda)P := BP_{\xi\xi} + \theta P_\xi + DF(u_0(\xi))P - \lambda P = 0, \quad P(\xi) \in \mathbb{C}^n, \tag{2.7}$$

can be taken as a second-order linear ordinary differential equation, which we write in a first-order system as follows:

$$Y' = A(\lambda, \xi)Y, \tag{2.8}$$

where

$$Y(\xi) = \begin{pmatrix} P(\xi) \\ Q(\xi) \end{pmatrix} \in \mathbb{C}^{2n}, \quad Q = P',$$

and

$$A(\lambda, \xi) = \begin{pmatrix} 0 & I \\ B^{-1}(\lambda I - DF(u_0(\xi))) & -\theta B^{-1} \end{pmatrix}.$$

For each λ, (2.8) has at least one nontrivial bounded solution up to multiplication of constants if and only if it is an eigenvalue of L, in which case the P component of such a solution is an eigenfunction of L.

We introduce a new variable τ via the relation

$$\xi = \frac{1}{\kappa} \log\left(\frac{1+\tau}{1-\tau}\right). \tag{2.9}$$

This turns (2.8) into an autonomous system on $\mathbb{C}^{2n} \times (-1, +1)$:

$$\begin{cases} Y' = A(\lambda, \tau)Y := A(\lambda, \xi(\tau))Y, \\ \tau' = \kappa(1 - \tau^2) \end{cases} \qquad (\,' = \tfrac{d}{d\xi}). \tag{2.10}$$

Letting

$$A(\lambda, \pm 1) = \begin{pmatrix} 0 & I \\ B^{-1}\left(\lambda I - DF\left(u_\pm\right)\right) & -\theta B^{-1} \end{pmatrix}, \tag{2.11}$$

we augment (2.10) into a system on $\mathbb{C}^{2n} \times [-1, +1]$, which is of class C^1 for sufficiently small $\kappa > 0$. This system is also denoted by

$$\begin{cases} Y' = A(\lambda, \tau)Y, \\ \tau' = \kappa(1 - \tau^2). \end{cases} \tag{2.12}$$

$(\mathbf{O}, -1)$ and $(\mathbf{O}, +1)$ are equilibria of (2.12). The next lemma shows that λ is an eigenvalue of L if and only if the unstable manifold $\mathbf{W}_-^u$ of $(\mathbf{O}, -1)$ and the stable manifold $\mathbf{W}_+^s$ of $(\mathbf{O}, +1)$ has a nontrivial intersection, namely, $\mathbf{W}_-^u \cap \mathbf{W}_+^s \neq \{\mathbf{O}\} \times (-1, 1)$.

Lemma 2.1. *Let $(Y(\xi), \tau(\xi))$ be a nontrivial solution of* (2.12).

1. *$Y(\xi)$ is bounded as $\xi \to -\infty$ if and only if $(Y(\xi), \tau(\xi)) \in \mathbf{W}_-^u$.*
2. *$Y(\xi)$ is bounded as $\xi \to +\infty$ if and only if $(Y(\xi), \tau(\xi)) \in \mathbf{W}_+^s$.*

Now we are at the final stage of constructing the stability index.

Proposition 2.2. *If $u_\pm$ are stable, then there is a constant $\beta < 0$ such that for all λ with* $\mathrm{Re}\lambda > \beta$, *$A(\lambda, -1)$ (resp. $A(\lambda, +1)$) have n eigenvalues with positive real parts and n with negative.*

This means that $\mathbf{W}_-^u\big|_{\mathbb{C}^{2n} \times \{\tau\}}$ and $\mathbf{W}_+^s\big|_{\mathbb{C}^{2n} \times \{\tau\}}$ are n-dimensional subspaces of $\mathbb{C}^{2n} \times \{\tau\}$. We regard them as n-dimensional vector bundles over intervals.

For each fixed λ, define a continuous map

$$\Phi_\lambda \colon [-1, 1) \to G_n(\mathbb{C}^{2n}) : \tau \mapsto \left[\mathbf{W}^u|_{\mathbb{C}^{2n} \times \{\tau\}}\right], \tag{2.13}$$

where $G_n(\mathbb{C}^{2n})$ is the Grassmannian manifold of n-dimensional subspaces in $\mathbb{C}^{2n}$.

Let $U_\pm^\lambda$ ($S_\pm^\lambda$) be the unstable (stable) subspace of $A(\lambda, \pm 1)$.

Lemma 2.2. *If λ is not an eigenvalue of L, then*

$$\Phi_\lambda(\tau) \to [U_+^\lambda] \quad (\tau \to +1) \tag{2.14}$$

locally uniformly in $G_n(\mathbb{C}^{2n})$ for $\lambda \notin \sigma(L)$.

Set

$$\mathfrak{K} := (K^\circ \times \{-1\}) \cup (K \times [-1, +1]) \cup (K^\circ \times \{+1\}) \cong \mathbb{S}^2 \tag{2.15}$$

and

$$\mathfrak{G} : \mathfrak{K} \to G_n(\mathbb{C}^{2n}) : \mathfrak{G}(\lambda, \tau) = \begin{cases} \left[U_-^\lambda\right], & \lambda \in K^\circ, \quad \tau = -1, \\ \Phi_\lambda(\tau), & \lambda \in K, \quad \tau \in (-1, 1), \\ \left[U_+^\lambda\right], & \lambda \in K^\circ, \quad \tau = +1. \end{cases} \tag{2.16}$$

By Lemma 2.2, $\mathfrak{G}$ is continuous and induces an n-dimensional vector bundle $\mathfrak{E}(K)$ over $\mathfrak{K}$.

Definition 2.2.

$$\mathfrak{E}(K) := \mathfrak{G}^*\big(\Gamma_n(\mathbb{C}^{2n})\big)\,, \tag{2.17}$$

where $\Gamma_n(\mathbb{C}^{2n})$ *is the canonical bundle over* $G_n(\mathbb{C}^{2n})$.

Alexander et al. showed that this bundle determines the stability of the traveling wave $u_0(\xi)$.

Theorem 2.1 (Alexander, Gardner, and Jones [1]). *The first Chern number* $c_1(\mathfrak{E}(K))$ *of the bundle* $\mathfrak{E}(K)$ *coincides with the number of eigenvalues of* L *in* K° *including their multiplicity.*

We modify this theorem slightly for later use.

Consider a system that (2.12) induces on $\overset{n}{\wedge}\mathbb{C}^{2n} \times [-1,+1]$,

$$\begin{cases} Y^{(n)\prime} = A^{(n)}(\lambda,\tau)Y^{(n)}, \\ \qquad \tau' = \kappa(1-\tau^2). \end{cases} \tag{2.18}$$

Here $Y^{(n)}(\xi) \in \overset{n}{\wedge}\mathbb{C}^{2n}$ and $A^{(n)} = A \otimes I \otimes \cdots \otimes I + \cdots + I \otimes I \otimes \cdots \otimes I \otimes A|_{\overset{n}{\wedge}\mathbb{C}^{2n}}$, where $\overset{n}{\wedge}\mathbb{C}^{2n}$ denotes the nth exterior power of $\mathbb{C}^{2n}$. Let $m = \binom{2n}{n} - 1$; then $\mathbb{CP}^m \cong \mathbb{P}(\overset{n}{\wedge}\mathbb{C}^{2n})$ and (2.18) defines a system of the following form on $\mathbb{CP}^m \times [-1,+1]$ since it is linear in the $Y^{(n)}$ component:

$$\begin{cases} z' = Z(z,\tau;\lambda), \\ \tau' = \kappa(1-\tau^2) \end{cases} \quad (z \in \mathbb{CP}^m). \tag{2.19}$$

Let $\tilde{U}^\lambda_\pm$ ($\tilde{S}^\lambda_\pm$) be points on $\mathbb{CP}^m$ corresponding to $\overset{n}{\wedge}U^\lambda_\pm$ ($\overset{n}{\wedge}S^\lambda_\pm$), i.e., $\tilde{U}^\lambda_\pm := \Pi(\overset{n}{\wedge}U^\lambda_\pm)$ (resp. $\tilde{S}^\lambda_\pm := \Pi(\overset{n}{\wedge}S^\lambda_\pm)$) for the projection $\Pi\colon \mathbb{C}^{m+1} \setminus \{\mathbf{O}\} \to \mathbb{CP}^m$; then $(\tilde{U}^\lambda_-, -1)$ and $(\tilde{U}^\lambda_+, +1)$ (resp. $(\tilde{S}^\lambda_-, -1)$ and $(\tilde{S}^\lambda_+, +1)$) are equilibria of (2.19). These equilibria are attractive in the invariant subspaces $\mathbb{CP}^m \times \{\tau = -1\}$ and $\mathbb{CP}^m \times \{\tau = +1\}$, which means that $(\tilde{U}^\lambda_-, -1)$ have a one-dimensional unstable direction in the whole space $\mathbb{CP}^m \times [-1,+1]$—namely, the τ direction—and $(\tilde{U}^\lambda_+, +1)$ is an attracting equilibrium. By Lemma 2.2, system (2.19) has a heteroclinic orbit from $(\tilde{U}^\lambda_-, -1)$ to $(\tilde{U}^\lambda_+, +1)$ for $\lambda \notin \sigma(L)$. This heteroclinic orbit is given by $\{(\tilde{\Phi}_\lambda(\tau), \tau)\}$, where $\tilde{\Phi}_\lambda(\tau) := \Pi(\overset{n}{\wedge}\Phi_\lambda(\tau))$. Define a map G by

$$G : \mathfrak{K} \to \mathbb{CP}^m : G(\lambda,\tau) = \begin{cases} \tilde{U}^\lambda_-, & \lambda \in K^\circ, \quad \tau = -1, \\ \tilde{\Phi}_\lambda(\tau), & \lambda \in K, \quad \tau \in (-1,1), \\ \tilde{U}^\lambda_+, & \lambda \in K^\circ, \quad \tau = +1; \end{cases} \tag{2.20}$$

then Theorem 2.1 can be restated as follows. (See Figure 2.1.)

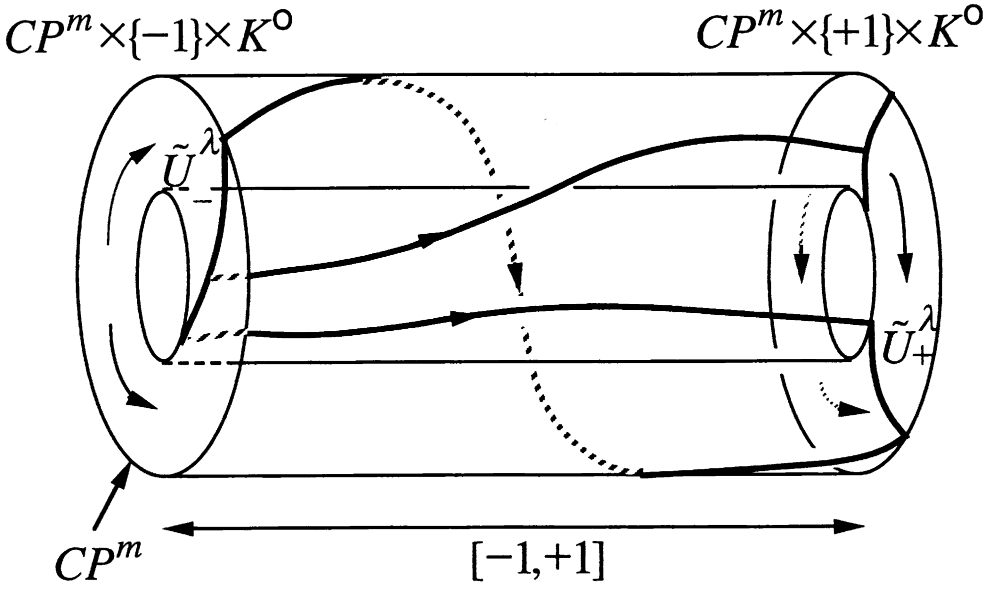
Theorem 2.2. $[G] \in \pi_2(\mathbb{CP}^m)) \cong \mathbb{Z}$ *counts the number of eigenvalues of* L *in* K°.

Proof. First, let us recall some parts of the proof of Theorem 2.1 in [1].

Let $H_- := (K^\circ \times \{-1\}) \cup (K \times I)$ and $H_+ := K^\circ \times \{+1\}$; then $H_- \cap H_+ = \mathbb{S}^1$. In [1], nonvanishing sections $\Gamma_-(\lambda,\tau)$ and $\Gamma_+(\lambda)$ of $\overset{n}{\wedge}\mathfrak{E}(K)$ over H_- and H_+ are constructed, i.e.,

(2.21)

$$\Gamma_- : H_- \to \overset{n}{\wedge}\mathbb{C}^{2n} : \Gamma_-(\lambda,\tau) \in \begin{cases} \Pi^{-1}(\tilde{U}^\lambda_-) = \overset{n}{\wedge}U^\lambda_-, & \lambda \in K^\circ, \quad \tau = -1, \\ \Pi^{-1}(\tilde{\Phi}_\lambda(\tau)) = \overset{n}{\wedge}\Phi_\lambda(\tau), & \lambda \in K, \quad \tau \in [-1,1], \end{cases}$$



$CP^m \times \{-1\} \times K^{\circ}$
$CP^m \times \{+1\} \times K^{\circ}$
$\tilde{U}_-^{\lambda}$
$\tilde{U}_+^{\lambda}$
CP^m
$[-1,+1]$

FIG. 2.1. *A schematic picture of Theorem* 2.2.

and

$$\Gamma_+ : H_+ \to \overset{n}{\wedge}\mathbb{C}^{2n} : \Gamma_+(\lambda) \in \Pi^{-1}(\tilde{U}_+^{\lambda}) = \overset{n}{\wedge} U_+^{\lambda}, \quad \lambda \in K^{\circ}, \tag{2.22}$$

where $\overset{n}{\wedge}\mathfrak{E}(K)$ is the exterior power of $\mathfrak{E}(K)$ and coincides with $G^*(\Gamma_1(\mathbb{C}^m))$. Trivializations of $\overset{n}{\wedge}\mathfrak{E}(K)$ over $H_\pm$ are defined with $\Gamma_\pm$ as

$$h_- : H_- \times \mathbb{C} \to \overset{n}{\wedge}\mathfrak{E}(K) : ((\lambda,\tau),z) \mapsto ((\lambda,\tau), z\Gamma_-(\lambda,\tau)), \tag{2.23}$$

$$h_+ : H_+ \times \mathbb{C} \to \overset{n}{\wedge}\mathfrak{E}(K) : ((\lambda,1),z) \mapsto ((\lambda,1), z\Gamma_+(\lambda)). \tag{2.24}$$

Then the map $g_E(\lambda)$ over $H_- \cap H_+$ defined by

$$h_+^{-1} \circ h_-((\lambda,1),z) = ((\lambda,1), g_E(\lambda)z) \tag{2.25}$$

can be seen as a map from $\mathbb{S}^1$ into $\mathbb{C}\setminus\{0\}$. In [1], it is shown that the homotopy class $[g_E] \in \pi_1(\mathbb{C}\setminus\{0\}) \cong \mathbb{Z}$ coincides with the number of eigenvalues of L in K°.

With $h_\pm$, we construct a map $h : \overset{n}{\wedge}\mathfrak{E}(K) \to \mathbb{C}^{m+1}$ as follows. Let N_- and N_+ be open neighborhoods of H_- and H_+ and extend $h_\pm$ over $N_\pm \times \mathbb{C}$. Urysone's lemma assures that there exist continuous functions $\alpha_\pm : N_\pm \to [0,1]$ satisfying $\alpha_\pm(\lambda,\tau) = 1$ if $(\lambda,\tau) \in H_\pm$ and $\alpha_\pm(\lambda,\tau) = 0$ if (λ,τ) is in some neighborhood of $\partial N_\pm$. Using this $\alpha_\pm$, define $h_i : \overset{n}{\wedge}\mathfrak{E}(K) \to \mathbb{C}$ $(i = 0,1)$ as

$$\begin{aligned} h_0((\lambda,\tau),v) &= \begin{cases} \alpha_-(\lambda,\tau)\cdot \pi_{h_-} \circ h_-^{-1}((\lambda,\tau),v), & (\lambda,\tau) \in N_-, \\ 0, & (\lambda,\tau) \notin N_-, \end{cases} \\ h_1((\lambda,\tau),v) &= \begin{cases} \alpha_+(\lambda,\tau)\cdot \pi_{h_+} \circ h_+^{-1}((\lambda,\tau),v), & (\lambda,\tau) \in N_+, \\ 0, & (\lambda,\tau) \notin N_+, \end{cases} \end{aligned} \tag{2.26}$$

where $\pi_{h_\pm} : H_\pm \times \mathbb{C} \to \mathbb{C}$ are projections. Then the map h is defined as

$$h : \overset{n}{\wedge}\mathfrak{E}(K) \to \mathbb{C}^{m+1} : h((\lambda,\tau),v) = (h_0((\lambda,\tau),v), h_1((\lambda,\tau),v), 0, \ldots, 0). \tag{2.27}$$

This h induces a map $\tilde{h} : \mathfrak{K} \to \mathbb{CP}^m$.

In what follows, we show that this $\tilde{h}$ is homotopic to G. First, let us assume that for any $((\lambda, \tau), v) \in \overset{n}{\wedge} \mathfrak{E}(K) \subset \mathfrak{K} \times \mathbb{C}^{m+1}$, $v \neq 0$, no negative real s satisfies $v = sh\left((\lambda, \tau), v\right)$. Under this condition,

$$h_s : \overset{n}{\wedge} \mathfrak{E}(K) \to \mathbb{C}^{m+1} : h_s((\lambda, \tau), v) := (1 - s)h((\lambda, \tau), v) + sv \tag{2.28}$$

induces a homotopy $\tilde{h}_s : \mathfrak{K} \to \mathbb{CP}^m$ with $\tilde{h}_0 = \tilde{h}$ and $\tilde{h}_1 = G$. When the assumption above is not satisfied, let $\mathbb{C}^{m+1}$ be identified with the subspace of $\mathbb{C}^{2m+2}$ as $\mathbb{C}^{m+1} \times \{0\} \subset \mathbb{C}^{m+1} \times \mathbb{C}^{m+1} = \mathbb{C}^{2m+2}$ and regard h as $h' : \overset{n}{\wedge} \mathfrak{E}(K) \to \mathbb{C}^{2m+2}$, and let h_{odd}, G', $G'_{\text{even}} : \overset{n}{\wedge} \mathfrak{E}(K) \to \mathbb{C}^{2m+2}$ be

$$\begin{aligned} h_{\text{odd}}((\lambda, \tau), v) &= (h_0((\lambda, \tau), v), 0, h_1((\lambda, \tau), v), 0, \ldots, 0), \\ G'((\lambda, \tau), v) &= (v, 0), \\ G'_{\text{even}}((\lambda, \tau), v) &= (0, v_0, 0, v_1, 0, \ldots, 0, v_m), \end{aligned} \tag{2.29}$$

where $v = (v_0, v_1, \ldots, v_m) \in \mathbb{C}^{m+1}$. Then h' and these maps induce the maps $\tilde{h}'$, $\tilde{h}_{\text{odd}}$, $\tilde{G}'$, $\tilde{G}'_{\text{even}} : \overset{n}{\wedge} \mathfrak{E}(K) \to \mathbb{CP}^{2m+1}$, and by the same argument as above,

$$\tilde{h}' \simeq \tilde{h}_{\text{odd}}, \qquad \tilde{h}_{\text{odd}} \simeq \tilde{G}'_{\text{even}}, \qquad \tilde{G}'_{\text{even}} \simeq \tilde{G}' \tag{2.30}$$

in $\mathbb{CP}^{2m+1}$. Thus $\tilde{h}'$ and $\tilde{G}'$ are homotopic in $\mathbb{CP}^{2m+1}$. Here $\mathbb{CP}^m$ can be regarded as a subspace of $\mathbb{CP}^{2m+1}$ corresponding to the inclusion $\mathbb{C}^{m+1} \hookrightarrow \mathbb{C}^{2m+2}$, and $\tilde{h} = \tilde{h}'$ and $G = \tilde{G}'$ by this identification. Thus $\tilde{h}$ is homotopic to G in $\mathbb{CP}^{2m+1}$.

Recall that $\mathbb{CP}^{2m+1}$ can be regarded as a cell complex as $\mathbb{CP}^{2m+1} = e^0 \cup e^2 \cup \cdots \cup e^{4m+2}$, where e^i $(i = 0, \ldots, 4m + 2)$ is an i-dimensional cell, and in this cell decomposition, $\mathbb{CP}^{2m+1} \setminus \mathbb{CP}^m \subset e^{m+1} \cup \cdots \cup e^{4m+2}$ consists of cells the dimension of which are equal or greater than 4. This means that the homotopy $\tilde{h} \simeq G$ can be realized in $\mathbb{CP}^m$ as $\tilde{h}, G : \mathfrak{K} \to \mathbb{CP}^m$ and $\mathfrak{K}$ is two dimensional. Thus $[\tilde{h}] = [G] \in \pi_2(\mathbb{CP}^m)$ with a natural identification corresponding to a change of base point.

The next step is to show that $[g_E] = [\tilde{h}]$ by a suitable isomorphism from $\pi_1(\mathbb{C}\setminus\{0\})$ to $\pi_2(\mathbb{CP}^m)$. The map $\tilde{h}$ can be regarded as $\tilde{h} : \mathfrak{K} \to \mathbb{CP}^1 \subset \mathbb{CP}^m$. Here again, the inclusion $\mathbb{CP}^1 \hookrightarrow \mathbb{CP}^m$ induces an isomorphism $\pi_2(\mathbb{CP}^1) \cong \pi_2(\mathbb{CP}^m)$, and by this homomorphism, $[\tilde{h}] \in \pi_2(\mathbb{CP}^m)$ is considered to be $[\tilde{h}] \in \pi_2(\mathbb{CP}^1)$. Recall that $h_+^{-1} \circ h_-((\lambda, 1), z) = ((\lambda, 1), g_E(\lambda)z)$, so if $\Gamma_-(\lambda, 1) = w(\lambda)\Gamma_+(\lambda)$ for $\lambda \in K$, $\tau = 1$, then $g_E(\lambda) = w(\lambda)$, i.e.,

$$g_E(\lambda) = \frac{h_1((\lambda, 1), v)}{h_0((\lambda, 1), v)}. \tag{2.31}$$

This means that g_E can be seen as

$$g_E : \mathbb{S}^1 \cong K \times \{+1\} \to \mathbb{CP}^1 \cong \mathbb{S}^2 : g_E(\lambda) = [h_0((\lambda, 1), v) : h_1((\lambda, 1), v)]. \tag{2.32}$$

Clearly, g_E is homotopic to some map $\bar{g} : \mathbb{S}^1 \to \mathbb{S}^1 = \big\{[1 : z] \,\big|\, |z| = 1\big\} \subset \mathbb{CP}^1$, and by suspension homomorphism, $[\bar{g}] \in \pi_1(\mathbb{S}^1)$ corresponds to $[\tilde{h}] \in \pi_2(\mathbb{CP}^1)$. Here this homomorphism is an isomorphism since $\pi_1(\mathbb{S}^1) \cong \pi_2(\mathbb{S}^2) \cong \mathbb{Z}$. Thus we get the desired result. □

Remark 2.2. We omit references to base points in the following argument because the isomorphism corresponding to a change of base points is uniquely determined by those points as $\pi_1\left(\mathbb{CP}^m\right) = 0$.

The whole arguments hold also for homology, but we employ homotopy for the sake of easier intuitive understanding.

3. An extension of the stability index.

In this section, we consider an extension of the stability index for perturbed systems and give a proof of one of our main theorems.

3.1. Perturbed systems. Consider a system of the form (2.2) with a parameter $\mu \in \mathbb{R}$:

$$u_t = Bu_{\xi\xi} + \theta(\mu)u_\xi + F(u;\mu). \tag{3.1}$$

Suppose that this system has a traveling wave $u_0(\xi)$ at $\mu = \mu_0$ that satisfies condition (T.W.). We assume without loss of generality that $F(u_\pm;\mu) = 0$ for all $\mu \in \mathbb{R}$. We also assume that $u_\pm$ are both stable for (2.1) or, in this case, for

$$u_t = Bu_{xx} + F(u;\mu). \tag{3.2}$$

We consider the equation of the stationary solutions for (3.1),

$$Bu_{\xi\xi} + \theta(\mu)u_\xi + F(u;\mu) = 0, \tag{3.3}$$

and we again write this as a first-order system,

$$\begin{cases} u' = v, \\ v' = -B^{-1}F(u;\mu) - \theta(\mu)B^{-1}v \end{cases} \qquad ('= \tfrac{d}{d\xi}). \tag{3.4}$$

Let $(u_0(\xi), u_0'(\xi))$ be the heteroclinic solution from $(u_-,0)$ to $(u_+,0)$ of (3.4) with $\mu = \mu_0$. Choose some point $(u_0^\dagger, v_0^\dagger)$ near $(u_+,0)$ on the orbit of $(u_0(\xi), u_0'(\xi))$, and take an n-dimensional plane Σ that intersects with the unstable manifold $\mathfrak{W}_-^u$ of $(u_-,0)$ transversely at $(u_0^\dagger, v_0^\dagger)$.

LEMMA 3.1. *When μ is near μ_0, $\mathfrak{W}_-^u \cap \Sigma = \big\{(u_\mu^\dagger, v_\mu^\dagger)\big\}$ (one point). Moreover, the point $(u_\mu^\dagger, v_\mu^\dagger)$ depends on μ continuously.* (*See Figure* 3.1.)

Proof. Consider (3.4) with a trivial equation $\mu' = 0$:

$$\begin{cases} u' = v, \\ v' = -B^{-1}F(u;\mu) - \theta(\mu)B^{-1}v, \\ \mu' = 0. \end{cases} \tag{3.5}$$

$S := \{(u_-,0,\mu) | \mu \in \mathbb{R}\}$ is a normally hyperbolic invariant set for (3.5), and let V^u be its unstable manifold. From the definition of Σ, V^u and $\Sigma \times \mathbb{R}$ intersect transversely at (u_0, v_0, μ_0) and, moreover, both V^u and $\Sigma \times \mathbb{R}$ are $(n+1)$-dimensional, so the intersection $V^u \cap \{\Sigma \times \mathbb{R}\}$ is a one-dimensional curve. The image under projection for (u,v) components is the curve $\{(u_\mu^\dagger, v_\mu^\dagger)\}$ parameterized by μ, and thus it is continuous. □

By translation on ξ, we can assume that $(u_0(0), u_0'(0)) = (u_0^\dagger, v_0^\dagger)$. We denote a solution that starts at $(u_\mu^\dagger, v_\mu^\dagger)$ when $\xi = 0$ by $(u(\xi;\mu), u'(\xi;\mu))$, namely, $(u(\xi;\mu_0), u'(\xi;\mu_0)) = (u_0(\xi), u_0'(\xi))$ and $u(\xi,\mu) \to u_-$ as $\xi \to -\infty$ since $(u_\mu^\dagger, v_\mu^\dagger) \in \mathfrak{W}_-^u$.

For this solution $u(\xi;\mu)$ and

$$\mathfrak{K}' = (K^\circ \times \{-1\}) \cup (K \times [-1,0]) \cong D^2,$$

we construct a map $G_-^\mu : \mathfrak{K}' \to \mathbb{CP}^m$ as follows.

Consider a system corresponding to (2.8),

$$Y' = A(\lambda, \xi; \mu)Y, \tag{3.6}$$

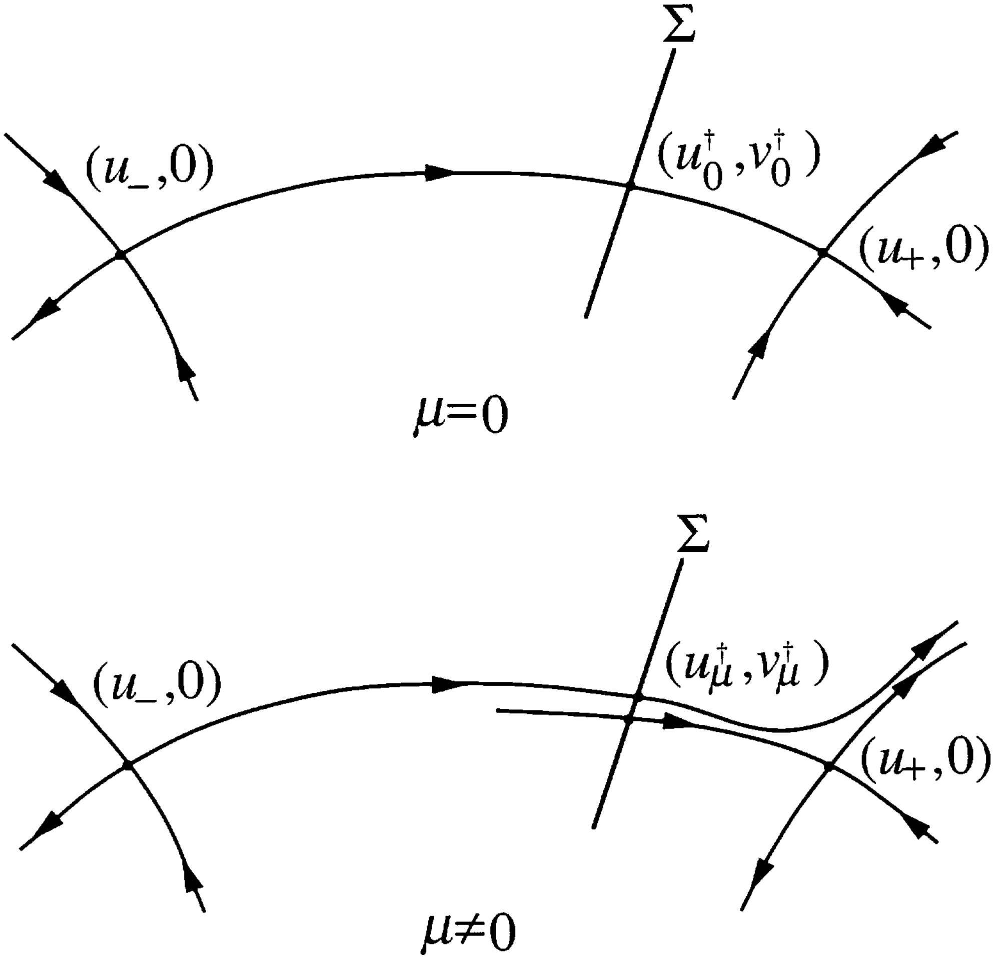

Σ
$(u_-,0)$
$(u_0^\dagger, v_0^\dagger)$
$(u_+,0)$
$\mu=0$
Σ
$(u_-,0)$
$(u_\mu^\dagger, v_\mu^\dagger)$
$(u_+,0)$
$\mu\neq 0$

FIG. 3.1. *The intersection point of* $\mathfrak{W}^u_-$ *and* Σ *for* μ *near* μ_0.

where

$$Y(\xi) = \begin{pmatrix} P(\xi) \\ Q(\xi) \end{pmatrix} \in \mathbb{C}^{2n}, \quad Q = P',$$

and

$$A(\lambda, \xi; \mu) = \begin{pmatrix} 0 & I \\ B^{-1}(\lambda I - DF(u(\xi;\mu))) & -\theta B^{-1} \end{pmatrix}.$$

A similar argument as before leads to a system on $\mathbb{C}^{2n} \times [-1, +1)$,

$$\begin{cases} Y' = A(\lambda, \tau; \mu) Y, \\ \tau' = \kappa(1 - \tau^2) \end{cases} \quad ({}' = \tfrac{d}{d\xi}), \tag{3.7}$$

where

$$A(\lambda, -1; \mu) = \begin{pmatrix} 0 & I \\ B^{-1}(\lambda I - DF(u_-)) & -\theta B^{-1} \end{pmatrix}.$$

Again, this equation induces a system on $\mathbb{CP}^m \times [-1, 1)$,

$$\begin{cases} z' = Z(z, \tau : \lambda; \mu), \\ \tau' = \kappa(1 - \tau^2) \end{cases} \qquad (z \in \mathbb{CP}^m). \tag{3.8}$$

We define a map G^{μ}_{-} as

$$G^{\mu}_{-} : \mathfrak{K}' \to \mathbb{CP}^m : G^{\mu}_{-}(\lambda, \tau) = \begin{cases} \tilde{U}^{\lambda}_{-}, & \lambda \in K^{\circ}, \quad \tau = -1, \\ \tilde{\Phi}^{\mu}_{\lambda}(\tau), & \lambda \in K, \quad \tau \in (-1, 0], \end{cases} \tag{3.9}$$

where $(\tilde{\Phi}^{\mu}_{\lambda}(\tau), \tau)$ is the unstable manifold of an equilibrium $(\tilde{U}^{\lambda}_{-}, -1)$ of (3.8). Obviously, this map is continuous and $G^{\mu_0}_{-} = G|_{\mathfrak{K}'}$ for G in (2.16).

Take a neighborhood N of $G((K \times [0, +1]) \cup (K^{\circ} \times \{+1\}))$ in $\mathbb{CP}^m$ satisfying $\pi_1(N) = 0$ and $\pi_2(N) = 0$. Such an N exists, for example, for K small and $(u^{\dagger}_0, v^{\dagger}_0)$ near $(u_+, 0)$.

Clearly, $G^{\mu}_{-}(\lambda, \tau)$ is also continuous in μ, so if μ is near μ_0, then

$$G^{\mu}_{-}\,(K \times \{0\}) \subset N. \tag{3.10}$$

This means that G^{μ}_{-} defines an element of the relative homotopy group $\pi_2\,(\mathbb{CP}^m, N)$.

Theorem 3.1.

$$\pi_2(\mathbb{CP}^m, N) \cong \pi_2(\mathbb{CP}^m) \cong \mathbb{Z},$$

and $[G^{\mu}_{-}] \in \pi_2(\mathbb{CP}^m, N)$ *corresponds to* $[G] \in \pi_2(\mathbb{CP}^m)$ *through the above isomorphism. Consequently,* $[G^{\mu}_{-}]$ *determines the stability of the travelling wave* $u_0(\xi)$ *at* $\mu = \mu_0$.

Proof. Consider the following exact sequence:

$$0 = \pi_2(N) \to \pi_2(\mathbb{CP}^m) \xrightarrow{p} \pi_2(\mathbb{CP}^m, N) \to \pi_1(N) = 0. \tag{3.11}$$

By the definition of N, $[G](\in \pi_2(\mathbb{CP}^m)$ corresponds to $[G^{\mu_0}_{-}]\,(\in \pi_2(\mathbb{CP}^m, N))$ through the homomorphism p. Moreover, (3.10) implies $[G^{\mu}_{-}] = [G^{\mu_0}_{-}]$ in $\pi_2\,(\mathbb{CP}^m, N)$. Here from the assumption on N, both ends of the above sequence are trivial, which implies that

$$\pi_2(\mathbb{CP}^m, N) \cong \pi_2(\mathbb{CP}^m) \cong \mathbb{Z},$$

and we get the desired result. □

Corollary 3.1. *The same argument holds for the perturbation of orbits. That is, at* $\mu = \mu_0$*, we can make a similar extension of the stability index for a solution of* (3.4) *on the unstable manifold* $\mathbf{W}^u_{-}$ *of* $(u_-, 0)$ *if it is near* $(u_0(\xi), u_0'(\xi))$. (*See Figure* 3.2.)

3.2. Additive formula for gluing bifurcation. In this section, we give a proof of an additive formula for the stability index under the gluing bifurcation of traveling waves. This theorem has been proven by Alexander and Jones [2] using cobordism invariance of the Chern number. Here we give another proof of additivity based on the above extension of the index.

Suppose system (3.1) has two traveling waves $u_1(\xi)$ and $u_2(\xi)$ when $\mu = \mu_0$ that satisfy

$$\begin{aligned} &u_1(\xi) \to u_- \quad (\xi \to -\infty), \qquad u_1(\xi) \to u_* \quad (\xi \to +\infty), \\ &u_2(\xi) \to u_* \quad (\xi \to -\infty), \qquad u_2(\xi) \to u_+ \quad (\xi \to +\infty), \end{aligned}$$

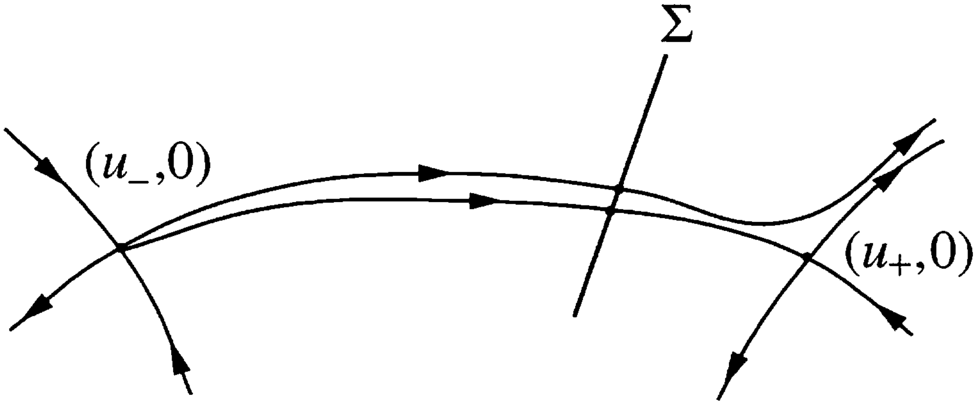


FIG. 3.2. *The case of the perturbation of an orbit.*

and (T.W.), where the stationary solutions $u_\pm$ and u_* are stable. Let $\Gamma_i = \{(u_i(\xi), u_i'(\xi))\}$ $(i = 1, 2)$ be the heteroclinic orbits of (3.4) corresponding to the traveling waves $u_i(\xi)$. Assume that system (3.4) undergoes a gluing bifurcation, that is, that the system has a heteroclinic orbit $\Gamma = \{(u(\xi), u'(\xi))\}$ from $(u_-, 0)$ to $(u_+, 0)$ for $\mu \approx \mu_0$ that stays in a tubular neighborhood of $\Gamma_1 \cup \Gamma_2$. (See Figure 3.3.)

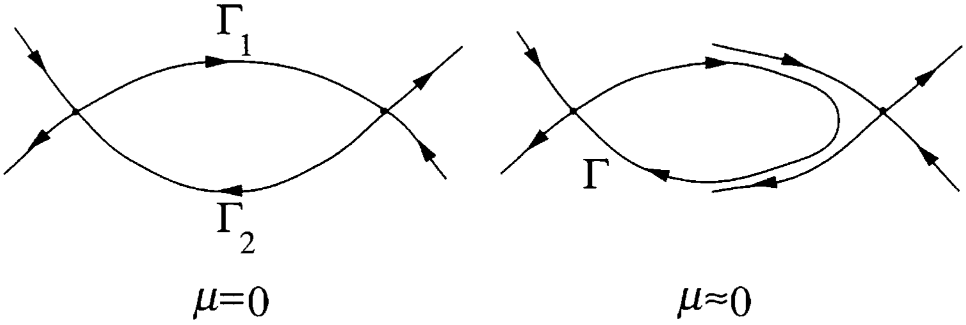


FIG. 3.3. *A gluing bifurcation.*

Fix a simple closed curve K on the complex plane $\mathbb{C}$, and let

$$G, G_i : \mathfrak{K} \to \mathbb{CP}^m, \quad i = 1, 2, \tag{3.12}$$

be the maps defined for $u(\xi)$ and $u_i(\xi)$.

In the situation above, we have the following additive formula.

THEOREM 3.2.

$$[G] = [G_1] + [G_2] \in \pi_2(\mathbb{CP}^m). \tag{3.13}$$

Proof. By definition, it holds that

$$G_1|_{(K\cup K^\circ)\times\{\tau=+1\}} = G_2|_{(K\cup K^\circ)\times\{\tau=-1\}} : K \cup K^\circ \to \mathbb{CP}^m. \tag{3.14}$$

Write this map as

$$H : K \cup K^\circ \to \mathbb{CP}^m. \tag{3.15}$$

First, let us assume that

$$\pi_1(H(K \cup K^\circ)) = 0 \quad \text{and} \quad \pi_2(H(K \cup K^\circ)) = 0. \tag{3.16}$$

Take a neighborhood N of $H(K \cup K^\circ)$ in $\mathbb{CP}^m$ that satisfies

$$\pi_1(N) = 0 \quad \text{and} \quad \pi_2(N) = 0. \tag{3.17}$$

If μ is close enough to μ_0 and as a result Γ is included in some small neighborhood of $\Gamma_1 \cup \Gamma_2$, we can assume that

$$\begin{cases} u(\xi) \approx u_1(\xi) & (\xi \leq 0), \\ u(\xi) \approx u_2(\xi) & (\xi \geq 0) \end{cases} \tag{3.18}$$

through some translations in the ξ variable. (See Figure 3.4.)

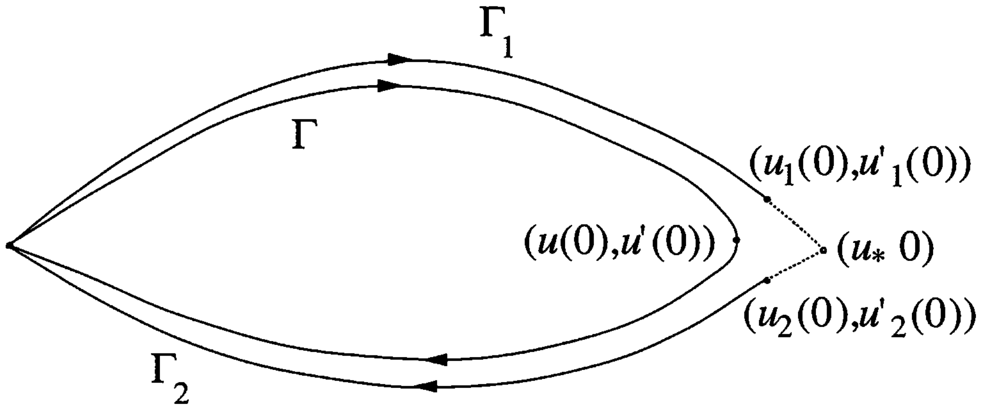


FIG. 3.4. Γ *and* $\Gamma_1 \cup \Gamma_2$.

This implies $u(0) \approx u_*$ and

$$G_1|_{K\times\{\tau=+1\}} \approx G_1|_{K\times\{\tau=0\}} \approx G|_{K\times\{\tau=0\}} \, : K \to \mathbb{CP}^m, \tag{3.19}$$

which means that

$$G(K \times \{0\}) \subset N. \tag{3.20}$$

Set

$$G_- := G|_{(K^\circ\times\{\tau=-1\})\cup(K\times[-1,0])}, \tag{3.21}$$

$$G_+ := G|_{(K\times[0,+1])\cup(K^\circ\times\{\tau=+1\})}, \tag{3.22}$$

$$G_{1-} := G_1|_{(K^\circ\times\{\tau=-1\})\cup(K\times[-1,+1])}, \tag{3.23}$$

$$G_{2+} := G_2|_{(K\times[-1,+1])\cup(K^\circ\times\{\tau=+1\})}; \tag{3.24}$$

then we have

$$G_- \simeq G_{1-} : (D^2, \partial D^2) \to (\mathbb{CP}^m, N), \tag{3.25}$$

$$G_+ \simeq G_{2+} : (D^2, \partial D^2) \to (\mathbb{CP}^m, N), \tag{3.26}$$

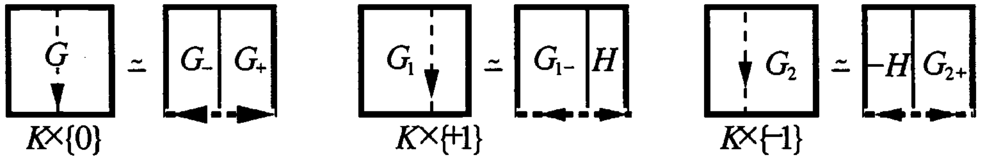


FIG. 3.5. *G, G_1, and G_2 as elements in $\pi_2(\mathbb{CP}^m, N)$.*

where $\simeq$ indicates homotopy.

Consider the following exact sequence again, where both ends are trivial:

$$0 = \pi_2(N) \to \pi_2(\mathbb{CP}^m) \xrightarrow{i} \pi_2(\mathbb{CP}^m, N) \to \pi_1(N) = 0. \tag{3.27}$$

From these definitions, we have (see Figure 3.5)

$$i([G]) = [G_-] + [G_+], \tag{3.28}$$

$$i([G_1]) = [G_{1-}] + [H], \tag{3.29}$$

$$i([G_2]) = -[H] + [G_{2+}]. \tag{3.30}$$

Here (3.25) and (3.26) imply that

$$[G_-] = [G_{1-}], \qquad [G_+] = [G_{2+}] \quad \text{in } \pi_2(\mathbb{CP}^m, N), \tag{3.31}$$

and by the definition of N, $[H] = 0$. This leads to

$$\begin{aligned} i\,([G]) &= [G_-] + [G_+] \\ &= [G_{1-}] + [G_{2+}] \\ &= i([G_1]) + i([G_2]) \quad \text{in } \pi_2(\mathbb{CP}^m, N). \end{aligned} \tag{3.32}$$

From this, we get the desired result, namely,

$$\begin{aligned} [G] &= i^{-1} \circ i([G]) \\ &= i^{-1} \circ i([G_1]) + i^{-1} \circ i\,([G_2]) \\ &= [G_1] + [G_2] \quad \text{in } \pi_2(\mathbb{CP}^m). \end{aligned} \tag{3.33}$$

For general K, divide K° into small parts; then each part satisfies condition (3.16) if the partition is fine enough, and therefore the above argument is applicable to each part. Collecting all of the parts, we get the result for K. □

4. Application.

4.1. Direct consequences of the additive formula. In this section, we consider some applications of Theorem 3.2. First, the next theorem is its direct consequence.

THEOREM 4.1. *Under the assumption of Theorem* 3.2, *$u(\xi)$ is unstable if either $u_1(\xi)$ or $u_2(\xi)$ is unstable.*

Proof. Suppose $u_1(\xi)$ is unstable, that is, the eigenvalue problem

$$(L_1 - \lambda)P := BP_{\xi\xi} + \theta P_\xi + DF(u_1(\xi))P - \lambda P \tag{4.1}$$

has an eigenvalue λ_1 with a positive real part. Take a simple closed curve $K_1 \subset \mathbb{C}$ satisfying

$$\lambda_1 \in K_1^\circ \quad \text{and} \quad K_1 \cup K_1^\circ \subset \{\lambda \in \mathbb{C} | \mathrm{Re}\lambda > 0\}. \tag{4.2}$$

By applying Theorem 3.2 to this K_1, we have an eigenvalue λ_0 inside K_1° of the eigenvalue problem

$$(L-\lambda)P := BP_{\xi\xi} + \theta P_\xi + DF(u(\xi))P - \lambda P, \tag{4.3}$$

which means that $\mathrm{Re}\lambda_0 > 0$, and hence $u(\xi)$ is unstable. (See Figure 4.1.)

The case where $u_2(\xi)$ is unstable can be treated similarly. □

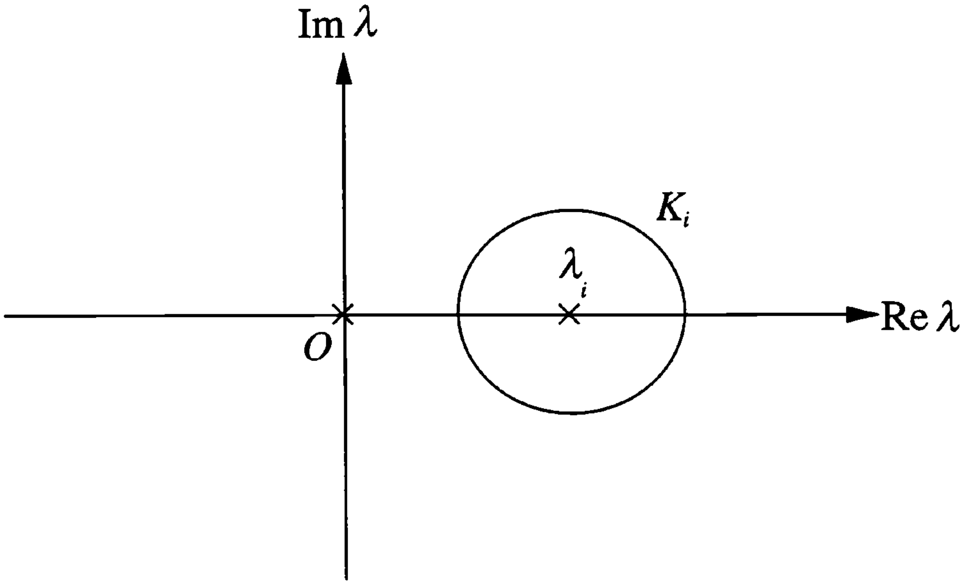


FIG. 4.1. *The eigenvalue with* $\mathrm{Re}\lambda_0 > 0$.

This theorem means that $u(\xi)$ can be stable only if both $u_1(\xi)$ and $u_2(\xi)$ are stable.

In what follows, let us suppose that both $u_1(\xi)$ and $u_2(\xi)$ are stable. In this case, Theorem 3.2 only tells us that L has two eigenvalues near the origin, and consequently we cannot decide the stability of $u(\xi)$ from this theorem alone. Actually, we have examples of both cases. (See section 5.)

4.2. Orientation index. Consider the eigenvalue problem for $\lambda \in \mathbb{R}$:

$$(L-\lambda)P := BP_{\xi\xi} + \theta P_\xi + DF(u_0(\xi))P - \lambda P = 0, \quad P(\xi) \in \mathbb{R}^n. \tag{4.4}$$

Again, we rewrite this as

$$\begin{cases} Y' = A(\lambda,\tau)Y, \\ \tau' = \kappa(1-\tau^2), \end{cases} \tag{4.5}$$

which defines a system on $\mathbb{RP}^m \times [-1,+1]$ as follows:

$$\begin{cases} z' = Z(z,\tau:\lambda), \\ \tau' = \kappa(1-\tau^2) \end{cases} \quad (z \in \mathbb{RP}^m). \tag{4.6}$$

Take $\lambda_1, \lambda_2 \in \mathbb{R}$ with $\beta < \lambda_1 < \lambda_2$, where β is as in Proposition 2.1, and assume that $\lambda_1, \lambda_2 \notin \sigma(L)$. Of course, Lemma 2.2 holds for $\lambda = \lambda_i$, $i = 1, 2$.

Set

$$\mathfrak{k} := ([\lambda_1, \lambda_2] \times \{-1\}) \cup (\{\lambda_1, \lambda_2\} \times [-1, +1]) \cup ([\lambda_1, \lambda_2] \times \{+1\}) \cong \mathbb{S}^1 \tag{4.7}$$

and let

$$g : \mathfrak{k} \to \mathbb{RP}^m \tag{4.8}$$

be defined in a similar manner as before, that is,

$$g(\lambda, \tau) = \begin{cases} \tilde{U}^{\lambda}_{-}, & \lambda \in [\lambda_1, \lambda_2], & \tau = -1, \\ \tilde{\Phi}_{\lambda}(\tau), & \lambda = \lambda_i,\ i = 1, 2, & \tau \in (-1, 1), \\ \tilde{U}^{\lambda}_{+}, & \lambda \in [\lambda_1, \lambda_2], & \tau = +1, \end{cases} \tag{4.9}$$

where $\tilde{U}^{\lambda}_{\pm}$ are the equilibria of (4.6) that correspond to the unstable subspaces $U^{\lambda}_{\pm}$ of $A(\lambda, \pm 1)$ and $(\tilde{\Phi}_{\lambda}(\tau), \tau)$ is a heteroclinic orbit connecting $(\tilde{U}^{\lambda}_{-}, -1)$ and $(\tilde{U}^{\lambda}_{+}, +1)$.

Next, for this g, we have as a part of Theorem 2.2.

THEOREM 4.2. *$[g] \in \pi_1(\mathbb{RP}^m) \cong \mathbb{Z}_2$ counts the parity of the number of eigenvalues of L on the interval $[\lambda_1, \lambda_2]$. That is, if the interval includes an odd number of eigenvalues, then $[g] = 1$, and if it includes an even number of eigenvalues, then $[g] = 0$.*

Proof. Theorem 2.1 was proven in [1] using the Evans function. First, we briefly summarize its construction.

Consider equations (2.18) and (2.19). An orbit $\{(Y(\lambda, \tau), \tau)\}_{-1<\tau<1}$ of (2.18) can be chosen so that $\Pi(Y(\lambda, \tau)) = \tilde{\Phi}_{\lambda}(\tau)$ for the projection $\Pi : \mathbb{C}^{m+1} \setminus \{\mathbf{O}\} \to \mathbb{CP}^m$ and for each τ, $Y(\lambda, \tau)$ is analytic in λ [1]. Similarly, an orbit $\{(Y^*(\lambda, \tau), \tau)\}_{-1<\tau<1}$ of (2.18) can be chosen so that $\Pi(Y^*(\lambda, \tau)) = \tilde{\Phi}^*_{\lambda}(\tau)$ and for each τ, $Y^*(\lambda, \tau)$ is analytic in λ, where $\{(\tilde{\Phi}^*_{\lambda}(\tau), \tau)\}_{-1<\tau<1}$ is the unique orbit of (2.19) satisfying

$$\lim_{\tau \to +1} (\tilde{\Phi}^*_{\lambda}(\tau), \tau) = (\tilde{S}^{\lambda}_{+}, +1). \tag{4.10}$$

Note that $Y(\lambda, \tau) \in \overset{n}{\wedge}(\mathbf{W}^u_{-}|_{\mathbb{C}^{2n} \times \{\tau\}})$ and $Y^*(\lambda, \tau) \in \overset{n}{\wedge}(\mathbf{W}^s_{+}|_{\mathbb{C}^{2n} \times \{\tau\}})$.

We define the Evans function $D(\lambda)$ as

$$D(\lambda) := Y(\lambda, 0) \wedge Y^*(\lambda, 0) \in \overset{2n}{\wedge}\mathbb{C}^{2n} \cong \mathbb{C}. \tag{4.11}$$

This function is analytic in λ, and the number of zeroes of D in K° coincides with the number of eigenvalues of L in K° including its multiplicity. For λ real, $Y(\lambda, \tau)$ and $Y^*(\lambda, \tau)$ can be taken so that they are both real analytic in λ and thus $D(\lambda)$ is real analytic.

Let $\Pi_s : \mathbb{R}^{m+1} \setminus \mathbf{O} \to \mathbb{S}^m$ and $\Pi_p : \mathbb{S}^m \to \mathbb{RP}^m$ be projections and set $\hat{\Phi}_{\lambda}(\tau) = \Pi_s(Y(\lambda, \tau))$ and $\hat{\Phi}^*_{\lambda}(\tau) = \Pi_s(Y^*(\lambda, \tau))$; then $\tilde{\Phi}_{\lambda}(\tau) = \Pi_p(\hat{\Phi}_{\lambda}(\tau))$ and $\tilde{\Phi}^*_{\lambda}(\tau) = \Pi_p(\hat{\Phi}^*_{\lambda}(\tau))$. We define $\hat{U}^{\lambda}_{\pm,i}$ and $\hat{S}^{\lambda}_{\pm,i}$ $(i = 1, 2)$ as $\Pi_p^{-1}(\tilde{U}^{\lambda}_{\pm}) = \{\hat{U}^{\lambda}_{\pm,1}, \hat{U}^{\lambda}_{\pm,2}\}$ and $\Pi_p^{-1}(\tilde{S}^{\lambda}_{\pm}) = \{\hat{S}^{\lambda}_{\pm,1}, \hat{S}^{\lambda}_{\pm,2}\}$ so that

$$\lim_{\tau \to \pm 1} \hat{\Phi}_{\lambda_1}(\tau) = \hat{U}^{\lambda_1}_{\pm,1} \quad \text{and} \quad \lim_{\tau \to \pm 1} \hat{\Phi}^*_{\lambda_1}(\tau) = \hat{S}^{\lambda_1}_{\pm,1} \tag{4.12}$$

hold. Moreover, $\hat{U}^{\lambda}_{\pm,i}$ and $\hat{S}^{\lambda}_{\pm,i}$ depend continuously on λ. Note that $\lim_{\tau \to +1} \hat{\Phi}^*_{\lambda_2}(\tau) = \hat{S}^{\lambda_2}_{+,1}$ holds by (4.10). If we consider $\hat{U}^{\lambda}_{\pm,i}, \hat{S}^{\lambda}_{\pm,i} \in \mathbb{R}^{m+1}$ by the inclusion

$\mathbb{S}^m \hookrightarrow \mathbb{R}^{m+1} = \overset{n}{\wedge}\mathbb{R}^{2n}$, the sign of $D(\lambda_1)$ coincides with that of $\hat{U}^{\lambda_1}_{+,1} \wedge \hat{S}^{\lambda_1}_{+,1}$ and thus with that of $\hat{U}^{\lambda_2}_{+,1} \wedge \hat{S}^{\lambda_2}_{+,1}$ by continuity. This means that $D(\lambda_1)D(\lambda_2) > 0$ if and only if

$$\lim_{\tau\to+1} \hat{\Phi}_{\lambda_2}(\tau) = \hat{U}^{\lambda_2}_{+,1} \tag{4.13}$$

and $D(\lambda_1)D(\lambda_2) < 0$ if and only if

$$\lim_{\tau\to+1} \hat{\Phi}_{\lambda_2}(\tau) = \hat{U}^{\lambda_2}_{+,2}. \tag{4.14}$$

In the former case, a lift $\hat{g} : \mathfrak{k} \to \mathbb{S}^m$ of $g : \mathfrak{k} \to \mathbb{RP}^m$ is expressed as

$$\hat{g}(\lambda,\tau) = \begin{cases} \hat{U}^{\lambda}_{-,1}, & \lambda \in [\lambda_1,\lambda_2], \qquad\quad \tau = -1, \\ \hat{\Phi}_{\lambda}(\tau), & \lambda = \lambda_i, \quad i = 1,2, \quad \tau \in (-1,1), \\ \hat{U}^{\lambda}_{+,1}, & \lambda \in [\lambda_1,\lambda_2], \qquad\quad \tau = +1, \end{cases} \tag{4.15}$$

and thus g is zero homotopic in $\mathbb{RP}^m$, whereas in the latter case, g cannot be lifted because $\lim_{\tau\to+1} \hat{\Phi}_{\lambda_2}(\tau) = \hat{U}^{\lambda_2}_{+,2} \neq \hat{U}^{\lambda_2}_{+,1}$, and thus g is not zero homotopic in $\mathbb{RP}^m$. This proves that $D(\lambda_1)D(\lambda_2) > 0$ if and only if $[g] = 0$ and that $D(\lambda_1)D(\lambda_2) < 0$ if and only if $[g] = 1$.

Here, since $D(\lambda)$ is a real analytic function on $[\lambda_1,\lambda_2]$, there are even zeros in $[\lambda_1,\lambda_2]$ if $D(\lambda_1)D(\lambda_2) > 0$ and odd zeros if $D(\lambda_1)D(\lambda_2) < 0$ including multiplicity. Thus the theorem holds. □

For $0 > \lambda_1$ close to 0 and λ_2 large, Alexander and Jones [2] showed that the direction of intersection of the unstable manifold of $(u_-,0)$ and the stable manifold of $(u_+,0)$ with respect to the traveling-wave speed θ determines the index $[g]$. More precisely, the sign of the derivative $\frac{dD}{d\lambda}|_{\lambda=0}$ of the Evans function $D(\lambda)$ at $\lambda = 0$ is determined as follows.

Assume that $u \equiv u_\pm$ are stationary solutions of the system and that for $\theta = \theta_0$ the system has a traveling wave. Let $X^{\theta_0}(\xi)$ be the corresponding heteroclinic solution from $(u_-,0)$ to $(u_+,0)$ at $\theta = \theta_0$ of

$$\begin{cases} u' = v, \\ v' = -B^{-1}F(u) - \theta B^{-1}v \end{cases} \qquad ({}' = \tfrac{d}{d\xi}). \tag{4.16}$$

Let $\mathfrak{W}^u_\theta$ be the unstable manifold of $(u_-,0)$ and $\mathfrak{W}^s_\theta$ be the stable manifold of $(u_+,0)$, both n-dimensional. Also, we assume that $\mathfrak{W}^u_\theta \cap \mathfrak{W}^s_\theta = \emptyset$, except for $\theta = \theta_0$. Extend $X^{\theta_0}(\xi)$ smoothly in θ to the solutions $X^\theta_u(\xi) \in \mathfrak{W}^u_\theta$ and $X^\theta_s(\xi) \in \mathfrak{W}^s_\theta$ for θ near θ_0.

For the traveling wave at $\theta = \theta_0$, define the solutions $(Y(\lambda,\tau),\tau)$ and $(Y^*(\lambda,\tau),\tau)$ of (2.18) so that $D(\lambda) > 0$ holds for large λ, and choose the vectors $V^1_u,\dots,V^{n-1}_u$ in $\mathbf{W}^u_-\big|_{\mathbb{C}^{2n}\times\{0\}}$ and $V^1_s,\dots,V^{n-1}_s$ in $\mathbf{W}^s_+\big|_{\mathbb{C}^{2n}\times\{0\}}$ so that

$$\begin{aligned} Y(0,0) &= X^{\theta_0\prime}(0) \wedge V^1_u \wedge \cdots \wedge V^{n-1}_u, \\ Y^*(0,0) &= X^{\theta_0\prime}(0) \wedge V^1_s \wedge \cdots \wedge V^{n-1}_s, \end{aligned}$$

which implies that $X^{\theta_0\prime}(0), V^1_u,\dots,V^{n-1}_u$ ($X^{\theta_0\prime}(0), V^1_s,\dots,V^{n-1}_s$) is a basis of $T_{X^{\theta_0}(0)}\mathfrak{W}^u_{\theta_0}$ (resp. $T_{X^{\theta_0}(0)}\mathfrak{W}^s_{\theta_0}$). Then we have

$$\begin{aligned} D(0) &= Y(0,0) \wedge Y^*(0,0) \\ &= \det\left(X^{\theta_0\prime}_u(0)V^1_u \cdots V^{n-1}_u X^{\theta_0\prime}_s(0)V^1_s \cdots V^{n-1}_s\right). \end{aligned}$$

Proposition 4.1 (Alexander and Jones [2]).

$$\frac{dD(\lambda)}{d\lambda}\bigg|_{\lambda=0} = \det\left(\left(\frac{dX_s^\theta(0)}{d\theta}\bigg|_{\theta=\theta_0} - \frac{dX_u^\theta(0)}{d\theta}\bigg|_{\theta=\theta_0}\right) V_u^1 \cdots V_u^{n-1} X^{\theta_0}{}'(0) V_s^1 \cdots V_s^{n-1}\right).$$

This proposition implies that the sign of $(\frac{dX_s^\theta(0)}{d\theta}|_{\theta=\theta_0} - \frac{dX_u^\theta(0)}{d\theta}|_{\theta=\theta_0})$ determines $[g]$ because if $\frac{dD}{d\lambda}(0) > 0$, then $D(\lambda_1) < 0$ for $\lambda_1 < 0$ close to 0, which means that $[g] = 1$ as $D(\lambda_2) > 0$ for λ_2 large, and if $\frac{dD}{d\lambda}(0) < 0$, then $D(\lambda_1) > 0$ for $\lambda_1 < 0$, so $[g] = 0$.

4.3. Gluing bifurcations and stability. In this section, we discuss the stability of a traveling-wave bifurcating from two traveling waves by a gluing bifurcation.

Consider an ordinary differential equation on $\mathbb{R}^{2n}$ depending on a parameter $\mu \in \mathbb{R}^k$ $(k \geq 2)$,

$$\dot{x} = f(x) + g(x;\mu) \quad (\dot{} = \tfrac{d}{dt}), \tag{4.17}$$

where f and g are smooth and $g(x;0) = 0$. Assume that (4.17) has three equilibria $\mathbf{O}_1(\mu)$, $\mathbf{O}_2(\mu)$, and $\mathbf{O}_3(\mu)$ and that the eigenvalues

$$-\eta_{n-1}^i(\mu), \dots, -\eta_1^i(\mu), -\rho^i(\mu), \nu^i(\mu), \kappa_1^i(\mu), \dots, \kappa_{n-1}^i(\mu)$$

of linearizations of F at each equilibrium satisfy

$$-\mathrm{Re}\eta_{n-1}^i(0) \leq \cdots \leq -\mathrm{Re}\eta_1^i(0) < -\rho^i(0) < 0 < \nu^i(0) < \mathrm{Re}\kappa_1^i(0) \leq \cdots \leq \mathrm{Re}\kappa_{n-1}^i(0).$$

Also, assume that for $\mu = 0$, the system

$$\dot{x} = f(x) \tag{4.18}$$

has heteroclinic orbits Γ_i of (4.18) from $\mathbf{O}_i(0)$ to $\mathbf{O}_{i+1}(0)$ $(i = 1, 2)$ simultaneously.

In what follows, we consider bifurcations of these heteroclinic orbits under the following nondegeneracy conditions.

1. For each i, the heteroclinic orbit $\Gamma_i = \{h_i(t)\}$ is tangent to the eigenspace associated with the eigenvalue $\nu^i(0)$ of linearization of f at $\mathbf{O}_i(0)$ as $t \to -\infty$ and the eigenspace associated with $-\rho^{i+1}(0)$ of $\mathbf{O}_{i+1}(0)$ as $t \to +\infty$.
2. For $\mu = 0$, the unstable manifold $\mathfrak{W}^u(\mathbf{O}_i)$ and the stable manifold $\mathfrak{W}^s(\mathbf{O}_{i+1})$ $(i = 1, 2)$ have a one-dimensional intersection, i.e., for all $p \in \Gamma_i$,

$$\dim(T_p\mathfrak{W}^u(\mathbf{O}_i) \cap T_p\mathfrak{W}^s(\mathbf{O}_{i+1})) = 1.$$

3. $\mathfrak{W}^u(\mathbf{O}_i(0))$ is transverse to the $(n+1)$-dimensional ν-stable manifold $\mathfrak{W}^{\nu,s}(\mathbf{O}_{i+1}(0))$ $(i = 1, 2)$ that is invariant and is tangent to the sum of the eigenspaces corresponding to $\nu^{i+1}(0)$, $-\rho^{i+1}(0)$, and $-\eta_j^{i+1}$ $(1 \leq j \leq n-1)$. Also, $\mathfrak{W}^s(\mathbf{O}_{i+1}(0))$ is transverse to the $(n+1)$-dimensional $(-\rho)$-unstable manifold corresponding to the eigenvalues $-\rho^i(0)$, $\nu^i(0)$, and $\kappa_k^i(0)$ $(1 \leq k \leq n-1)$.
4. For a nontrivial bounded solution $\hat{q}^i(t)$ $(i = 1, 2)$ of the linear ordinary differential equation

$$\dot{\hat{z}} = -{}^tDf(h_i(t))\hat{z} \quad (i = 1, 2), \tag{4.19}$$

the vectors given by the integrals

$$\int_{-\infty}^{+\infty} \hat{q}^i(s) \cdot \frac{\partial}{\partial\mu} g(h_i(s); 0) ds \tag{4.20}$$

are linearly independent and hence nonzero.

Remark 4.1. The bounded solution $\hat{q}^i(t)$ is unique up to multiplication by constants.

Under these conditions, we have the following result.

PROPOSITION 4.2 (Kokubu [17]). *Under conditions* 1–4 *above, there exist two hypersurfaces M_i $(i = 1, 2)$ of codimension* 1 *in a sufficiently small neighborhood of $\mu = 0$ in $\mathbb{R}^k$ so that each of M_i consists of parameter values μ for which the system has a heteroclinic orbit Γ_i. Moreover, M_1 and M_2 intersect transversely at $\mu = 0$.*

In order to investigate the existence of a heteroclinic orbit from $\mathbf{O}_1$ to $\mathbf{O}_3$, we divide our analysis into two cases:

(i) $\nu^2(0) \neq \rho^2(0)$ (the case of noncritical eigenvalues);

(ii) $\nu^2(0) = \rho^2(0)$ (the case of critical eigenvalues).

PROPOSITION 4.3 (Kokubu [17]). *Under conditions* 1–4 *and for the case of noncritical eigenvalues, there exists a hypersurface M of codimension* 1 *with the boundary*

$$\partial M = M_1 \cap M_2$$

in a sufficiently small neighborhood of $\mu = 0$ in $\mathbb{R}^k$ so that M consists of parameter values μ for which the system has a heteroclinic orbit $\Gamma = \{h(t)\}$ from O_1 to O_3. Moreover,

(a) *if $\nu^2(0) < \rho^2(0)$, then M is tangent to M_2 at $\mu = 0$, and*

(b) *if $\nu^2(0) > \rho^2(0)$, then M is tangent to M_1 at $\mu = 0$.*

For the case of critical eigenvalues, we impose a further condition as follows:

5. The set $\{\mu | \nu^2(\mu) = \rho^2(\mu)\}$ forms a surface Π in the parameter space $\mathbb{R}^k$ and is transverse to both M_1 and M_2 at $\mu = 0$.

PROPOSITION 4.4 (Kokubu [17]). *Under conditions* 1–5 *and for the case of critical eigenvalues, there exists a hypersurface M of codimension* 1 *with the boundary*

$$\partial M = M_1 \cap M_2$$

in a sufficiently small neighborhood of $\mu = 0$ in $\mathbb{R}^k$ so that M consists of parameter values μ for which the system has a heteroclinic orbit $\Gamma = \{h(t)\}$ from O_1 to O_3. Moreover, M is tangent to neither M_1 nor M_2 at $\mu = 0$ in Π.

The proofs of Propositions 4.3 and 4.4 are as follows.

We may assume that $\nu^2 < \rho^2$, for otherwise we can apply the same argument by replacing t with $-t$. Take the heteroclinic solution $h_i(t)$ so that $h_i(0)$ is near $\mathbf{O}_2$, and let Σ_i be a properly chosen $(2n-1)$-dimensional plane that passes through $h_i(0)$ and is transverse to $\dot{h}_i(0)$. Let $q^i(t)$ be a solution of the variational equation

$$\dot{z} = Df(h_i(t))z \tag{4.21}$$

along $h_i(t)$, for which the next limits exist and are nonzero, and let $q^1(t)$ point to $\mathbf{O}_3$ along the heteroclinic orbit Γ_2 in the limit of $t \to +\infty$ and $q^2(t)$ point to $\mathbf{O}_1$ along the heteroclinic orbit Γ_1 in the limit of $t \to -\infty$:

$$\lim_{t \to -\infty} |q^i(t)| e^{\rho^i(0)t}, \qquad \lim_{t \to +\infty} |q^i(t)| e^{-\nu^{i+1}(0)t}. \tag{4.22}$$

We remark that the adjoint $\hat{q}^i(t)$ of $q^i(t)$ is a nontrivial bounded solution of equation (4.19) for each i. In this situation, there are two points $x_1^u(\mu) \in \mathfrak{W}^u(\mathbf{O}_1) \cap \Sigma_1$ and $x_2^s(\mu) \in \mathfrak{W}^s(\mathbf{O}_2) \cap \Sigma_1$ depending on μ smoothly, and their difference can be expressed as

$$x_1^u - x_2^s = \alpha(\mu) q^1(0); \tag{4.23}$$

similarly, there are two points $x_2^u(\mu) \in \mathfrak{W}^u(\mathbf{O}_2) \cap \Sigma_2$ and $x_3^s(\mu) \in \mathfrak{W}^s(\mathbf{O}_3) \cap \Sigma_2$ such that

$$x_3^s - x_2^u = \beta(\mu) q^2(0). \tag{4.24}$$

The function $\alpha(\mu)$ ($\beta(\mu)$) measures the separation of $\mathfrak{W}^u(\mathbf{O}_1)$ and $\mathfrak{W}^s(\mathbf{O}_2)$ in Σ_1 ($\mathfrak{W}^u(\mathbf{O}_2)$ and $\mathfrak{W}^s(\mathbf{O}_3)$ in Σ_2). For $\alpha(\mu) > 0$ and $\beta(\mu) > 0$, $\mathfrak{W}^u(\mathbf{O}_1)$ intersects with Σ_2, and we can also find a point $\tilde{x}_1^u(\mu) \in \mathfrak{W}^u(\mathbf{O}_1) \cap \Sigma_2$ such that

$$x_3^s - \tilde{x}_1^u = \chi(\mu) q^2(0), \tag{4.25}$$

where $\chi(\mu)$ measures the separation of $\mathfrak{W}^u(\mathbf{O}_1)$ and $\mathfrak{W}^s(\mathbf{O}_3)$ in Σ_2. $\chi(\mu)$ can be extended to the parameter values for which $\alpha(\mu) = 0$ or $\beta(\mu) = 0$. Proposition 4.2 was proven by showing that

$$\frac{d}{d\mu}\alpha(0) = |q^1(0)| \int_{-\infty}^{+\infty} \hat{q}^1(s) \frac{\partial}{\partial \mu} g(h_1(s); 0) ds \tag{4.26}$$

and

$$\frac{d}{d\mu}\beta(0) = |q^2(0)| \int_{-\infty}^{+\infty} \hat{q}^2(s) \frac{\partial}{\partial \mu} g(h_2(s); 0) ds, \tag{4.27}$$

where the right-hand sides of the equations are nonzero and linearly independent. Proposition 4.3 is proved by showing that $\chi(\mu) > 0$ if $\alpha(\mu) = 0$ and $\beta(\mu) > 0$, $\chi(\mu) < 0$ if $\alpha(\mu) > 0$ and $\beta(\mu) = 0$, and $\frac{d}{d\mu}\chi(0) = \frac{d}{d\mu}\beta(0)$.

For the proof of the existence part of Proposition 4.4, take $a = \alpha(\mu)$, $b = \beta(\mu)$, and $\lambda = \rho^2(\mu)/\nu^2(\mu) - 1$ as parameters and consider in the parameter space (a, b, λ), where $\lambda \geq 0$. Then the separation $\chi = \chi(a, b, \lambda)$ is defined in this parameter space, smooth in b and λ, and Lipschitz continuous in a. For the derivative of χ with respect to b, we have $\frac{\partial \chi}{\partial b}(0) = 1$. By a modification of the implicit-function theorem, we get a unique solution $b = b(a, \lambda)$ of $\chi(a, b, \lambda) = 0$ that is smooth in λ and Lipschitz in a. A similar argument for $\lambda \leq 0$ gives a unique solution $a = a(b, \lambda)$ of $\chi(a, b, \lambda) = 0$, and these coincide for $\lambda = 0$.

For some reaction-diffusion equations, this kind of bifurcation takes place. (See the examples below.) In such cases, from the stability of the traveling waves corresponding to the heteroclinic orbits Γ_1 and Γ_2, with some information about the twisting of these orbits, we can determine the stability of the traveling wave corresponding to the heteroclinic orbit Γ. We explain this in what follows.

First, we determine $(\frac{dX_s^\theta(0)}{d\theta}|_{\theta=\theta_0} - \frac{dX_u^\theta(0)}{d\theta}|_{\theta=\theta_0})$ in Proposition 4.1 from the above conditions. For the case where $\nu^2 < \rho^2$, take $X_u^\mu(\xi) \in \mathfrak{W}_\mu^u(\mathbf{O}_1)$ of Proposition 4.1 as $X_u^\mu(0) = \tilde{x}_1^u(\mu)$ and $X_s^\mu(\xi) \in \mathfrak{W}_s^\mu(\mathbf{O}_3)$ as $X_s^\mu(0) = x_3^s(\mu)$ so that

$$X_s^\mu(0) - X_u^\mu(0) = \chi(\mu) q^2(0). \tag{4.28}$$

Then

$$\left.\frac{\partial X_s^\mu(0)}{\partial \mu}\right|_{\mu=0} - \left.\frac{\partial X_u^\mu(0)}{\partial \mu}\right|_{\mu=0} = \frac{\partial \chi}{\partial \mu}(0) q^2(0) = \frac{\partial \beta}{\partial \mu}(0) q^2(0). \tag{4.29}$$

This means that if μ includes the traveling-wave speed θ or its translation $\theta - \theta_0$ as one of the parameters, then $(\frac{dX_s^\theta(0)}{d\theta}|_{\theta=\theta_0} - \frac{dX_u^\theta(0)}{d\theta}|_{\theta=\theta_0})$ is near $\frac{\partial \beta}{\partial \theta}(0) q^2(0)$. In particular, the sign of

$$\det\left(\left(\left.\frac{dX_s^\theta(0)}{d\theta}\right|_{\theta=\theta_0}-\left.\frac{dX_u^\theta(0)}{d\theta}\right|_{\theta=\theta_0}\right)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)$$

coincides with that of

$$\det\left(\frac{\partial\beta}{\partial\theta}(0)q^2(0)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)$$

for $\mu\approx 0$ if $\frac{\partial\beta}{\partial\theta}(0)\neq 0$.

For the case where $\nu^2=\rho^2$, for $\alpha>0$ and $\beta>0$, χ is also smooth in α, although it is not necessarily smooth in α at $\mu=0$. Because $\frac{\partial\chi}{\partial\beta}=1$, $d\chi(\mu)\neq 0$ for $\mu\in M$ if $\mu\approx 0$, $\alpha>0$, and $\beta>0$, which means that $\frac{\partial\chi}{\partial\theta}\neq 0$ if the θ direction is transverse to M. From this we can easily conclude that if the sign of $\frac{\partial\chi}{\partial\theta}$ coincides with that of $\frac{\partial\chi}{\partial\beta}$ at some $\mu\in M$, then the sign of

$$\det\left(\left(\left.\frac{dX_s^\theta(0)}{d\theta}\right|_{\theta=\theta_0}-\left.\frac{dX_u^\theta(0)}{d\theta}\right|_{\theta=\theta_0}\right)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)$$

coincides with the sign of

$$\det\left(\frac{\partial\beta}{\partial\theta}(0)q^2(0)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)$$

for the same μ, and vice versa. We remark that $\chi(\mu)<0$ for μ near the α-axis and $\chi(\mu)>0$ for μ near the β-axis, so we know the sign of $\frac{\partial\chi}{\partial\theta}$ from the bifurcation diagram. We state the above argument as a theorem.

THEOREM 4.3. *Assume that system* (3.4) *associated with the existence problem of traveling waves for system* (2.1) *undergoes the bifurcation as described in Propositions* 4.3 *and* 4.4. *Then if $\nu^2<\rho^2$, we have the following equality for the derivative of the Evans function that is defined for the traveling wave corresponding to the heteroclinic orbit* Γ:

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right)=\operatorname{sign}\left(\det\left(\frac{\partial\beta}{\partial\theta}(0)q^2(0)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)\right).$$

The same equality holds for $\nu^2=\rho^2$ if the sign of $\frac{\partial\chi}{\partial\theta}$ coincides with that of $\frac{\partial\chi}{\partial\beta}$, and if it differs, then

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right)=-\operatorname{sign}\left(\det\left(\frac{\partial\beta}{\partial\theta}(0)q^2(0)V_u^1\cdots V_u^{n-1}{X^{\theta_0}}'(0)V_s^1\cdots V_s^{n-1}\right)\right).$$

We also obtain a similar result for $\nu^2\geq\rho^2$ by replacing $\xi(=x-\theta t)$ with $-\xi$ in system (3.4).

Of course, Theorem 4.3 is not enough to determine the stability. We must know the orientation of $(V_u^1,\dots,V_u^{n-1})$ or, equivalently, that of $(V_s^1,\dots,V_s^{n-1})$ to determine the sign of $\frac{dD}{d\lambda}(0)$. To do this, we have to use other properties of the system.

In what follows, let us assume that system (3.4) undergoes the bifurcation as in Propositions 4.3 and 4.4 with $\mathbf{O}_1=\mathbf{O}_3$. In such a case, the two heteroclinic orbits Γ_1 and Γ_2 for $\mu=0$ form a loop called a heteroclinic loop and the orbit $\Gamma=\{h(\xi)|\xi\in\mathbb{R}\}$

bifurcating from the loop is a homoclinic orbit. In this situation, the orientation of $(V_u^1, \dots, V_u^{n-1})$ with respect to the homoclinic orbit is easily determined as follows.

Let $(\hat{V}_{u,1}^0, \dots, \hat{V}_{u,1}^{n-1})$ (resp. $(\hat{V}_{s,1}^0, \dots, \hat{V}_{s,1}^{n-1})$) be a basis of the unstable (resp. stable) eigenspace of $\mathbf{O}_1$ satisfying

$$\hat{V}_{u,1}^0 = \lim_{\xi \to -\infty} \frac{h_1(\xi)}{|h_1(\xi)|}, \qquad \hat{V}_{s,1}^0 = \lim_{\xi \to +\infty} \frac{h_2(\xi)}{|h_2(\xi)|},$$

and

$$\det \left(\hat{V}_{u,1}^0 \cdots \hat{V}_{u,1}^{n-1} \hat{V}_{s,1}^0 \cdots \hat{V}_{s,1}^{n-1} \right) > 0.$$

Then $(\hat{V}_{u,1}^0, \dots, \hat{V}_{u,1}^{n-1})$ determines the orientation of the local unstable manifold of $\mathbf{O}_1$ and propagates with the orientation of the global unstable manifold $\mathfrak{W}^u(\mathbf{O}_1)$ of $\mathbf{O}_1$. Let $(\hat{V}_{u,2}^0, \dots, \hat{V}_{u,2}^{n-1})$ be a basis of the unstable eigenspace of $\mathbf{O}_2$ with

$$\hat{V}_{u,2}^0 = \lim_{\xi \to -\infty} \frac{h_2(\xi)}{|h_2(\xi)|},$$

and let

$$\hat{V}_{s,2}^0 = \lim_{\xi \to +\infty} \frac{h_1(\xi)}{|h_1(\xi)|}.$$

From assumption 2 for the heteroclinic orbit Γ_1, the tangent space $T_{h_1(\xi)}\mathfrak{W}^u(\mathbf{O}_1)$ is tangent to the space spanned by $\hat{V}_{s,2}^0, \hat{V}_{u,2}^1, \dots, \hat{V}_{u,2}^{n-1}$ in the limit of $\xi \to +\infty$. We determine the orientation of $(\hat{V}_{u,2}^1, \dots, \hat{V}_{u,2}^{n-1})$ so that the orientation of $(\hat{V}_{s,2}^0, \hat{V}_{u,2}^1, \dots, \hat{V}_{u,2}^{n-1})$ is compatible with that of $\mathfrak{W}^u(\mathbf{O}_1)$, and then the orientation of $\mathfrak{W}^u(\mathbf{O}_2)$ is naturally determined by that of $(\hat{V}_{u,2}^0, \hat{V}_{u,2}^1, \dots, \hat{V}_{u,2}^{n-1})$. Similarly, the orientation of $\mathfrak{W}^u(\mathbf{O}_2)$ again determines the orientation of the unstable eigenspace of $\mathbf{O}_1$, but this orientation is not necessarily compatible with that of $(\hat{V}_{u,1}^0, \dots, \hat{V}_{u,1}^{n-1})$, which we defined earlier.

Definition 4.1. *The heteroclinic loop that consists of* Γ_1 *and* Γ_2 *is said to be nontwisted with respect to the strong unstable direction if the above orientation is compatible with that of* $(\hat{V}_{u,1}^0, \dots, \hat{V}_{u,1}^{n-1})$ *defined at the beginning. Otherwise, it is said to be twisted with respect to the strong unstable direction.*

Twisting with respect to the strong stable direction is similarly defined.

We have another definition of twisting, which is directly related to the structure of bifurcation. (See, for example, Deng [5] and the references therein.)

Consider $q^2(\xi)$, which is the solution of the variational equation (4.21) along $h_2(\xi)$ given in the proof of Propositions 4.3 and 4.4. From the requirement for $q^2(\xi)$ (see (4.22)), we may assume that

$$\lim_{\xi \to -\infty} q^2(\xi) e^{\rho^2 \xi} = -\hat{V}_{s,2}^0$$

and

$$\lim_{\xi \to +\infty} q^2(\xi) e^{-\nu^1 \xi} = c\hat{V}_{u,1}^0$$

for some nonzero constant c.

Definition 4.2. *The heteroclinic orbit Γ_2 is nontwisted if c is positive and twisted if c is negative.*

With these two kinds of twistings, we define the sign σ of $\Gamma = \{h(\xi)\}$ as follows.

Definition 4.3. $\sigma = +1$ *if either of the following holds.*

(1) *The heteroclinic loop is nontwisted with respect to the strong unstable direction and Γ_2 is nontwisted.*

(2) *The heteroclinic loop is twisted with respect to the strong unstable direction and Γ_2 is twisted.*

Otherwise, $\sigma = -1$.

Lemma 4.1. *If either of the following is satisfied for the heteroclinic loop that consists of two heteroclinic orbits Γ_1 and Γ_2, then $\sigma = +1$; otherwise, $\sigma = -1$.*

(1) *The heteroclinic loop is nontwisted with respect to the strong stable direction and Γ_1 is nontwisted.*

(2) *The heteroclinic loop is twisted with respect to the strong stable direction and Γ_1 is twisted.*

Proof. By the definitions of the two kinds of twistings above, the heteroclinic loop consisting of two heteroclinic orbits Γ_1 and Γ_2 that is nontwisted (twisted) with respect to the strong unstable direction is nontwisted (twisted) with respect to the strong stable direction if both heteroclinic orbits Γ_1 and Γ_2 are nontwisted or both are twisted. Otherwise, the loop that is nontwisted with respect to the strong unstable direction is twisted with respect to the strong stable direction, and vice versa. The lemma clearly holds. □

Then we can determine the sign of $\frac{dD}{d\lambda}(0)$.

Theorem 4.4.

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right) = \sigma \cdot \operatorname{sign}\left(\frac{\partial\beta}{\partial\theta}\right)$$

if $\nu^2 < \rho^2$. The same equality holds for $\nu^2 = \rho^2$ if the sign of $\frac{\partial\chi}{\partial\theta}$ coincides with that of $\frac{\partial\chi}{\partial\beta}$, and if it differs, then

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right) = -\sigma \cdot \operatorname{sign}\left(\frac{\partial\beta}{\partial\theta}\right).$$

For the case where $\nu^2 > \rho^2$, the following holds:

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right) = -\sigma \cdot \operatorname{sign}\left(\frac{\partial\alpha}{\partial\theta}\right).$$

The same equality holds for $\nu^2 = \rho^2$ if the sign of $\frac{\partial\chi}{\partial\theta}$ coincides with that of $\frac{\partial\chi}{\partial\alpha}$ for χ of the time-reversed system, and if it differs, then

$$\operatorname{sign}\left(\left.\frac{dD(\lambda)}{d\lambda}\right|_{\lambda=0}\right) = \sigma \cdot \operatorname{sign}\left(\frac{\partial\alpha}{\partial\theta}\right).$$

Proof. Define the orientation of the unstable manifold $\mathfrak{W}^u(\mathbf{O}_i)$ $(i = 1, 2)$ and the stable manifold $\mathfrak{W}^s(\mathbf{O}_i)$ of $\mathbf{O}_i$ continuously depending on μ by

$$\hat{V}^0_{u,1} = \lim_{\xi\to-\infty} \frac{h(\xi)}{|h(\xi)|}$$

and

$$\hat{V}^0_{s,1} = \lim_{\xi\to+\infty} \frac{h(\xi)}{|h(\xi)|}.$$

Let $V^1_u(\xi), \ldots, V^{n-1}_u(\xi)$ $(V^1_u(\xi), \ldots, V^{n-1}_u(\xi))$ be solutions of the variational equation along $h(\xi)$ such that $h'(\xi), V^1_u(\xi), \ldots, V^{n-1}_u(\xi)$ is a properly ordered basis of $T_{h(\xi)}\mathfrak{W}^u(\mathbf{O}_1)$ and $h'(\xi), V^1_s(\xi), \ldots, V^{n-1}_s(\xi)$ is a basis of $T_{h(\xi)}\mathfrak{W}^s(\mathbf{O}_1)$. Then the orientation of $\mathfrak{W}^u(\mathbf{O}_1)$ propagates along $h(\xi)$, which approaches to $\mathbf{O}_2$ once and comes back to $\mathbf{O}_1$. By definition, the orientation of $\mathfrak{W}^u(\mathbf{O}_1)$ coincides with that of $\mathfrak{W}^u(\mathbf{O}_2)$ when $h(\xi)$ is near $\mathbf{O}_2$. This implies that the orientation of $(\hat{V}^0_{u,1},\ V^1_u(\xi), \ldots, V^{n-1}_u(\xi))$ is compatible with that of $(\hat{V}^0_{u,1},\ \hat{V}^1_{u,1}, \ldots, \hat{V}^{n-1}_{u,1})$ for large ξ if and only if the heteroclinic loop is nontwisted. Therefore, the sign of

$$\det\left(\hat{V}^0_{u,1} V^1_u(\xi) \cdots V^{n-1}_u(\xi) h'(\xi) V^1_s(\xi) \cdots V^{n-1}_s(\xi)\right)$$

is positive for large ξ if the heteroclinic loop is nontwisted and negative if the heteroclinic loop is twisted. Therefore, by the definition of twisting of Γ_2, it follows that

$$\operatorname{sign}\left(\det\left(q^2(\xi)\ V^1_u(\xi) \cdots V^{n-1}_u(\xi) h'(\xi) V^1_s(\xi) \cdots V^{n-1}_s(\xi)\right)\right)$$

is positive if condition (1) or (2) of Definition 4.3 holds and negative otherwise. This proves the theorem.

By replacing ξ with $-\xi$ and taking into account Lemma 4.1 and the definition of χ (see (4.25)) in the proof of Propositions 4.3 and 4.4, we get the result for the case where $\nu^2 > \rho^2$. □

Remark 4.2.

(1) The heteroclinic loop is nontwisted with respect to the strong direction if system (3.4) has a certain symmetry for $\mu = 0$. For example, $h_2(\xi) = -h_1(\xi)$ or $h_2(\xi) = (h^u_2(\xi), h^v_2(\xi)) = (h^u_1(-\xi), -h^v_1(-\xi))$, where h^u_i denotes the u-component of h_i and h^v_i denotes the v-component of h_i. Note that a standing wave (a traveling wave with velocity $\theta = 0$) possesses the latter type of symmetry.

(2) When the dimension of system (3.4) is equal to or less than 3, the twisting of the loop is determined only by the twistings of Γ_1 and Γ_2.

Theorem 4.4 means that if σ is given, then we can determine the stability of the traveling wave corresponding to the homoclinic orbit $h(\xi)$ only from the bifurcation diagram. We summarize this in Figure 5.7.

5. Examples. In this section, we consider two examples of reaction-diffusion equations that possesses traveling waves forming a heteroclinic loop, and we study the stability of bifurcating waves by applying the theorems of this paper.

Example 1. We treat the following activator–inhibitor system:

$$\begin{cases} \varepsilon\tau u_t = \varepsilon^2 u_{xx} + f(u, v; \gamma, \nu), \\ \quad v_t = v_{xx} + g(u, v; \gamma, \nu), \end{cases} \tag{5.1}$$

where ε and τ are real positive parameters, ε is sufficiently small, and

$$\begin{cases} f(u, v; \gamma, \nu) = -u^3 + u - v, \\ g(u, v; \gamma, \nu) = u - \gamma v + \nu. \end{cases} \tag{5.2}$$

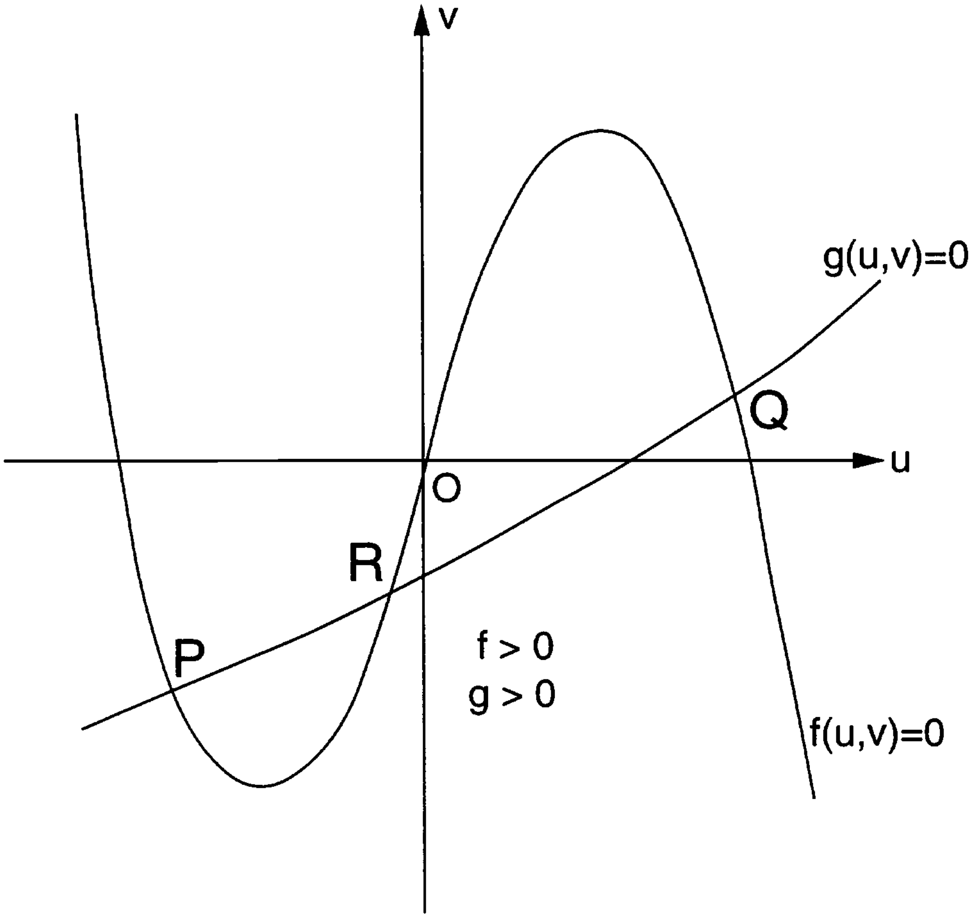


FIG. 5.1. *Nullclines of f and g in Example* 1.

The nullcline of f intersects with that of g at three points P, Q, and R, where P and Q are stable and R is an unstable constant solution of (5.1) (see Figure 5.1).

In what follows, we introduce the traveling coordinate $\xi = x - \theta t$, and we regard γ, ν, and the wave speed θ as bifurcation parameters. Consider

$$\text{(5.3)} \qquad \begin{cases} u' = \frac{1}{\varepsilon} u_1, \\ v' = v_1, \\ u_1' = -\frac{\theta\tau}{\varepsilon} u_1 - \frac{1}{\varepsilon} f(u, v; \gamma, \nu), \\ v_1' = -\theta v_1 - g(u, v; \gamma, \nu) \end{cases} \qquad ({}' = \tfrac{d}{d\xi})$$

corresponding to the existence of a traveling wave for (5.1). This system possesses heteroclinic orbits $\Gamma_1 = \{h_1(\xi; \gamma, \nu)\}$ from $(P, \mathbf{O})$ to $(Q, \mathbf{O})$ and $\Gamma_2 = \{h_2(\xi; \gamma, \nu)\}$ from $(Q, \mathbf{O})$ to $(P, \mathbf{O})$. The traveling (standing if $\theta = 0$) wave corresponding to Γ_1 is called a traveling (standing) front and Γ_2 is called a traveling (standing) back. Figure 5.2 depicts the situation in the parameter space, where M_1 corresponds to Γ_1 and M_2 corresponds to Γ_2. Notice that the thick curve shows the coexistence of both heteroclinic orbits or, in other words, the existence of what is called a heteroclinic loop.

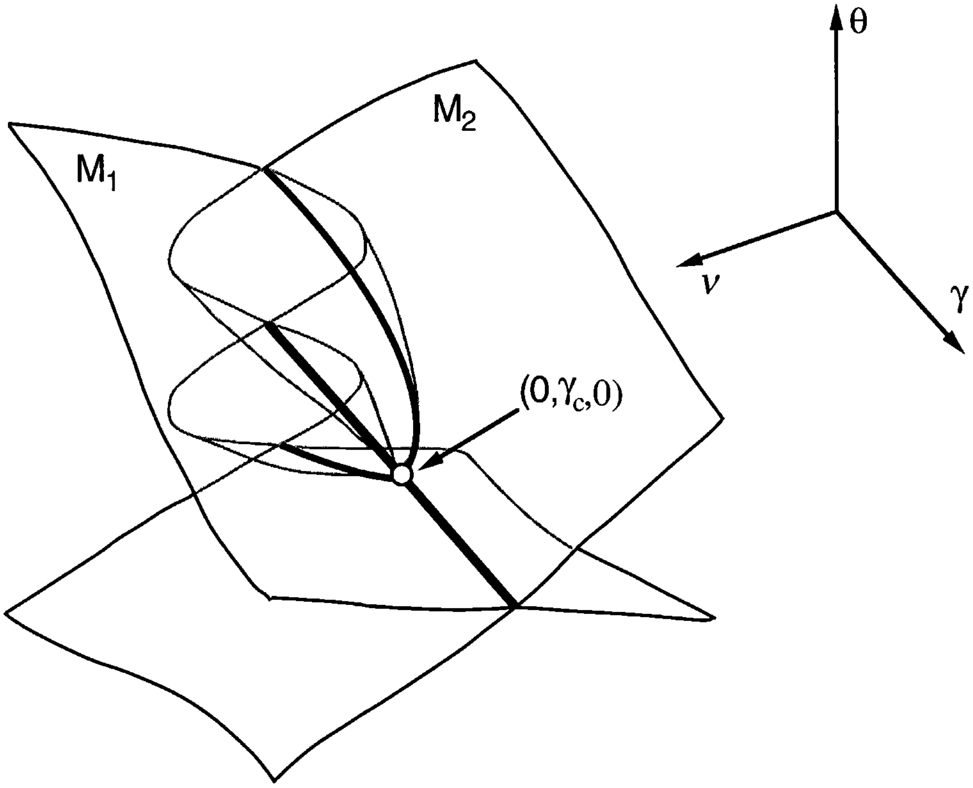


FIG. 5.2. *Bifurcation diagram of the traveling front and traveling back in Example* 1.

In Figure 5.3, bifurcation diagrams for ν and θ for fixed γ are shown.

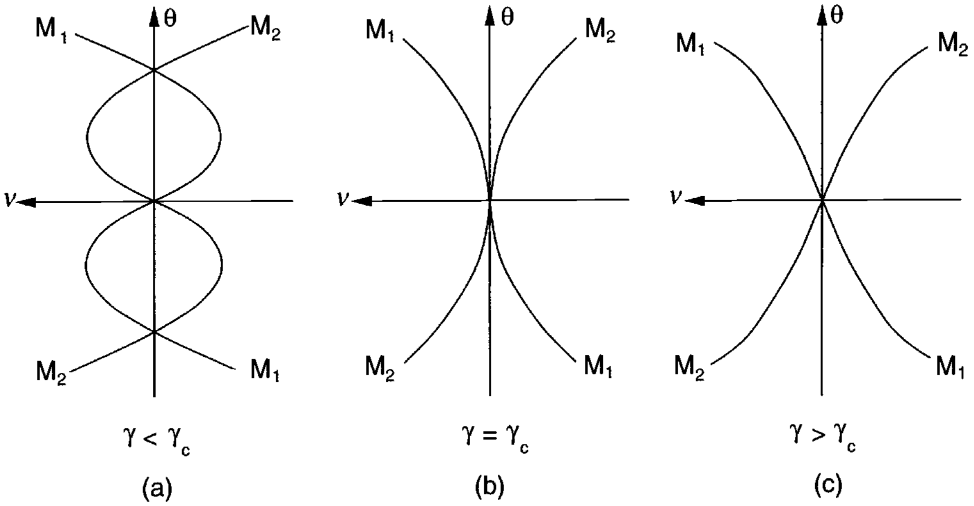


FIG. 5.3. *Bifurcation diagram of the traveling front and traveling back in Example* 1 *for fixed* γ.

Kokubu, Nishiura, and Oka [18] showed the existence of homoclinic orbits $\Gamma_P = \{h_P(\xi)\}$ associated with $(P, \mathbf{O})$ and $\Gamma_Q = \{h_Q(\xi)\}$ associated with $(Q, \mathbf{O})$ bifurcating from the heteroclinic loops; these homoclinic orbits correspond to traveling (standing) pulses. We show bifurcation diagrams in Figures 5.4 and 5.5, where M_P and M_Q correspond to these waves.

The stability and instability of fronts and backs are known. Ikeda [11] proved the stability and instability of pulses in [18] using the singular-perturbation method. In this example, we prove the stability and instability, applying the theorems in this paper without using the information in the singular limit.

First, note that the standing front and back that correspond to the heteroclinic loop for $(\gamma, \nu, \theta) = (\gamma, 0, 0)$ and $\gamma < \gamma_c$ are unstable. From Theorem 4.1, both standing pulses associated with P and Q bifurcating from the heteroclinic loop are unstable. Fronts and backs corresponding to the heteroclinic loops for other parameter values are known to be stable, and thus the pulses other than the above can be stable, and this is indeed the case. In such cases, the eigenvalue problem (2.7) for pulses has one eigenvalue near the origin other than the one at the origin, and it is real. We show that the derivative of the Evans function at the origin $\frac{dD}{d\lambda}(0)$ is positive, which means that $[g]$ in Theorem 4.2 is equal to 1 for $\lambda_1 < 0$ near 0 and λ_2 large, i.e., 0 is the only eigenvalue whose real part is equal or grater than 0. Thus the pulse is stable.

The heteroclinic orbits Γ_1 and Γ_2 are known to be nontwisted for $\nu = 0$. Notice that system (5.3) has the symmetry

$$h_2(\xi; \gamma, 0) = -h_1(\xi; \gamma, 0) \tag{5.4}$$

for the heteroclinic loop, and so it is nontwisted with respect to the strong direction (Remark 4.2). From Theorem 4.4 and the bifurcation diagram (Figure 5.5), we have

$$\operatorname{sign}\left(\frac{dD}{d\lambda}(0)\right) > 0,$$

and therefore the pulses are stable.

Example 2. Our next example is a Lotka–Volterra competition system:

$$\begin{cases} \quad u_t = u_{\xi\xi} + f(u,v), \\ \quad v_t = dv_{\xi\xi} + g(u,v), \qquad \xi \in \mathbb{R}, \quad t > 0, \\ u(0,\xi) \geq 0, \qquad v(0,\xi) \geq 0, \end{cases} \tag{5.5}$$

where

$$\begin{cases} f(u,v) = (1-u-cv)u, \\ g(u,v) = (a-bu-v)v, \end{cases} \tag{5.6}$$

and $a, b, c,$ and d are positive constants.

This system has two stable constant solutions $(u,v) \equiv (0,a)$ and $(u,v) \equiv (1,0)$ for $1/c < a < b$. Again consider the following system in search of traveling waves:

$$\begin{cases} u' = u_1, \\ v' = v_1, \\ u_1' = -\theta u_1 - (1-u-cv)u, \\ v_1' = -\frac{\theta}{d} v_1 - \frac{1}{d}(a-bu-v)v \end{cases} \qquad ({}' = \tfrac{d}{d\xi}). \tag{5.7}$$

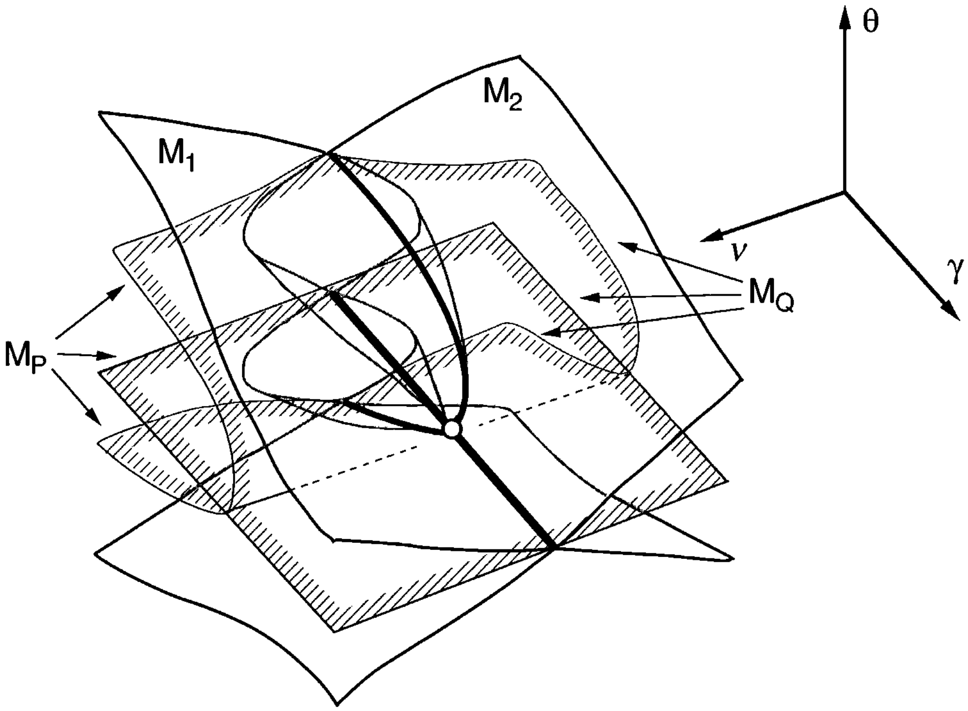


FIG. 5.4. *Bifurcation diagram of the traveling pulses in Example* 1.

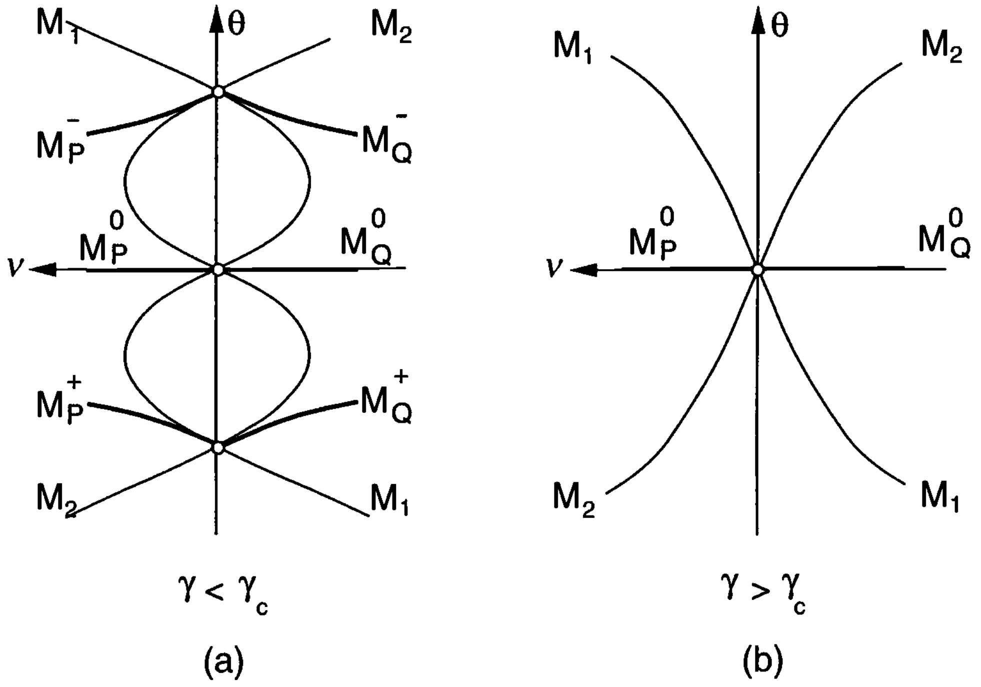


FIG. 5.5. *Bifurcation diagram of the traveling pulses in Example* 1 *for fixed* γ.

For each fixed d, there exist C^1 families $h_1(\xi; a, b, c)$ and $\theta(a, b, c)$ defined on $\mathfrak{P} = \{(a, b, c)|0 < 1/c < a < b\}$ such that h_1 is a solution of (5.7) with $\theta = \theta(a, b, c)$ satisfying

$$\lim_{\xi\to-\infty} h_1(\xi; a, b, c) = (0, a, 0, 0), \qquad \lim_{\xi\to+\infty} h_1(\xi : a, b, c) = (1, 0, 0, 0), \tag{5.8}$$

and $\theta_a(a, b, c) > 0$, $\theta_b(a, b, c) < 0$, and $\theta_c(a, b, c) > 0$. Moreover, there exists a C^1 family $a = a(b, c) \in (1/c, b)$ on $\{(b, c)|0 < 1/c < b\}$ such that $\theta(a(b, c), b, c) = 0$ holds (see Kan-on [13]). $\theta = 0$ means that this system also has symmetry, by which we can get a heteroclinic solution $h_2(\xi; a, b, c)$ satisfying

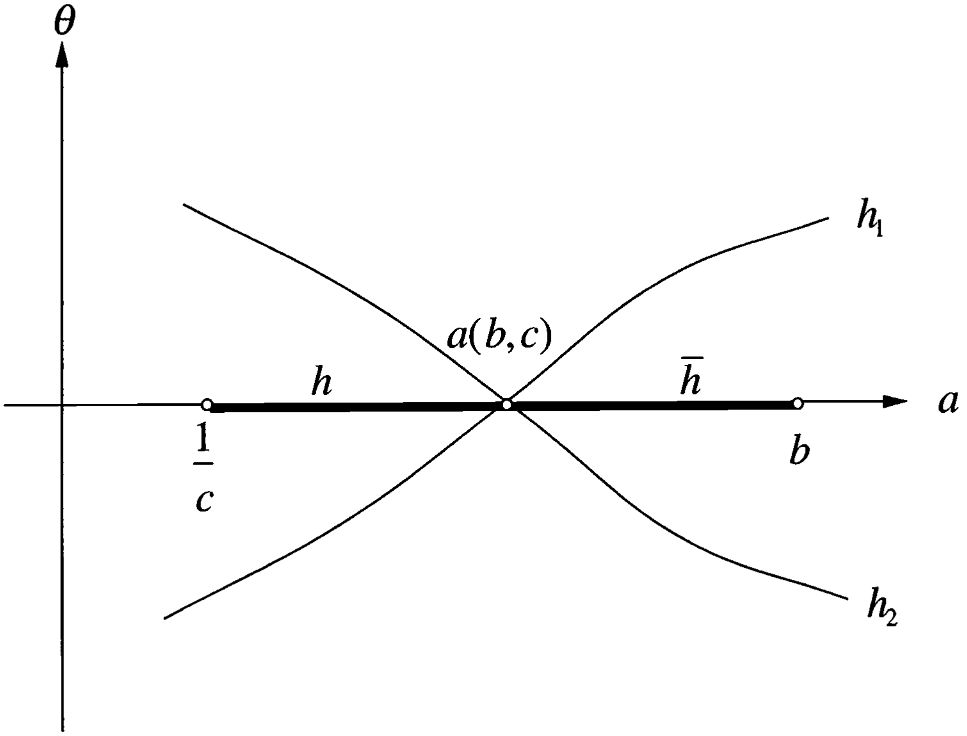


FIG. 5.6. *Bifurcation diagram in Example* 2.

$$\lim_{\xi\to-\infty} h_2(\xi; a, b, c) = (1, 0, 0, 0), \qquad \lim_{\xi\to+\infty} h_2(\xi : a, b, c) = (0, a, 0, 0) \tag{5.9}$$

by setting

(5.10)
$$h_2(\xi; a, b, c) = (h_1^1(-\xi; a, b, c), h_1^2(-\xi; a, b, c), -h_1^3(-\xi; a, b, c), -h_1^4(-\xi; a, b, c)),$$

where

$$h_1(\xi; a, b, c) = (h_1^1(\xi; a, b, c), h_1^2(\xi; a, b, c), h_1^3(\xi; a, b, c), h_1^4(\xi; a, b, c)). \tag{5.11}$$

This means that for $(\theta, a, b, c) = (0, a(b, c), b, c)$, the system possesses a nontwisted heteroclinic loop consisting of h_1 and h_2, and it is known that both h_1 and h_2 are

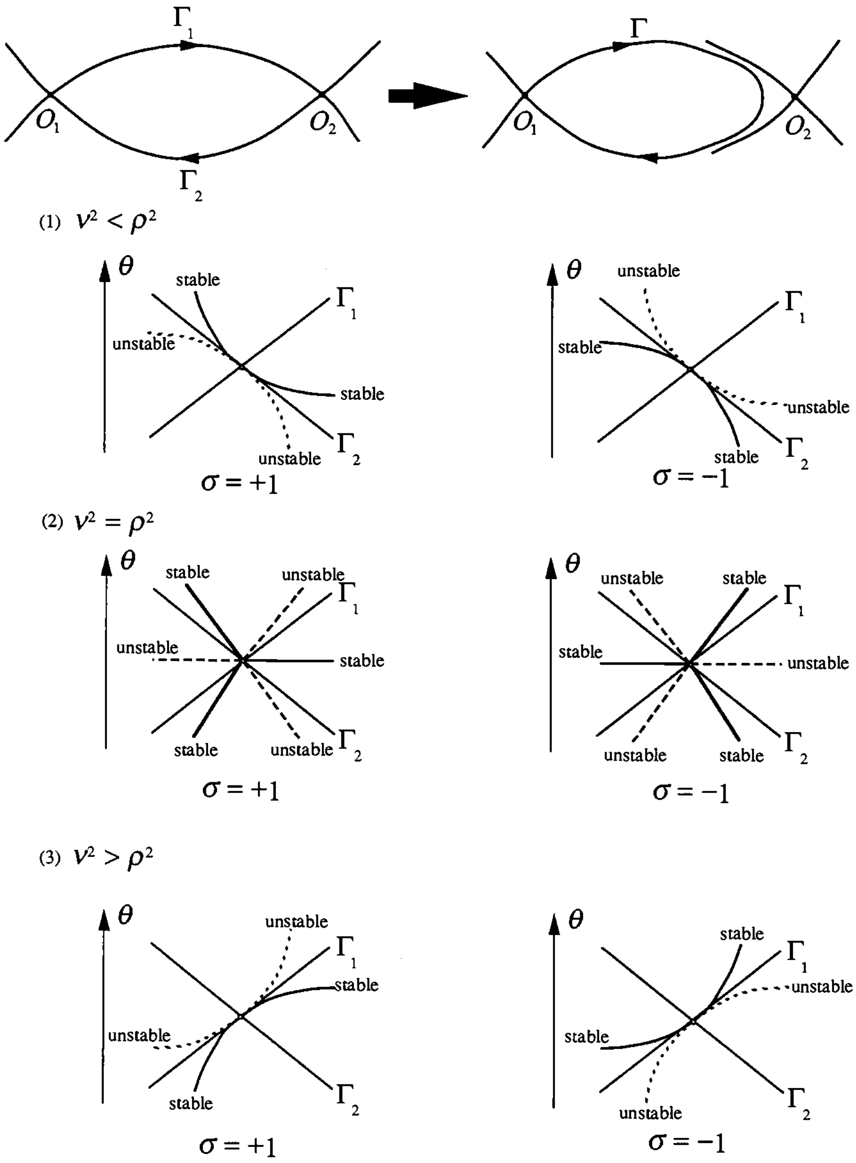


FIG. 5.7. *Stability of bifurcating pulses.*

nontwisted. Of course, system (5.7) has homoclinic orbits bifurcating from the heteroclinic loop (see Kan-on [14]). More precisely, there exists a C^1 family of solutions of (5.7) $h(\xi; a, b, c)$ ($\bar{h}(\xi; a, b, c)$) with $\theta = 0$ defined on $\mathfrak{Q} = \{(a, b, c) | 0 < 1/c < b, 1/c < a < a(b, c)\}$ ($\bar{\mathfrak{Q}} = \{(a, b, c) | 0 < 1/c < b, a(b, c) < a < b\}$) that satisfies

$$(5.12) \qquad \lim_{\xi \to \pm\infty} h(\xi; a, b, c) = (0, a, 0, 0) \qquad \left(\lim_{\xi \to \pm\infty} \bar{h}(\xi; a, b, c) = (1, 0, 0, 0) \right).$$

Again, traveling waves corresponding to the heteroclinic orbits are stable [16]. Kan-on [15] proved the instability of pulses by constructing an eigenfunction that corresponds to a positive eigenvalue.

However, by applying Theorem 4.4, we readily see from the bifurcation diagram (Figure 5.6) that the pulses are unstable.

Acknowledgments. The author expresses his sincere gratitude to Professors T. Nishida and H. Kokubu for valuable advice and encouragement. The author is also grateful to Professors H. Ikeda, C. Jones, Y. Kan-on, Y. Nishiura, and E. Yanagida for their useful comments and stimulating discussions.

REFERENCES

[1] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J. Alexander and C. Jones, *Existence and stability of asymptotically oscillatory double pulses*, J. Reine Angew. Math., 446 (1994), pp. 49–79.

[3] J. Alexander and C. Jones, *Existence and stability of asymptotically oscillatory triple pulses*, Z. Angew. Math. Phys., 44 (1993), pp. 189–200.

[4] P. Bates and C. Jones, *Invariant manifolds for semilinear partial differential equations*, Dynam. Report., 2 (1988), pp. 1–38.

[5] B. Deng, *The bifurcations of countable connections from a twisted heteroclinic loop*, SIAM J. Math. Anal., 22 (1991), pp. 653–678.

[6] J. Evans, *Nerve axon equations* III: *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577–594.

[7] J. Evans, *Nerve axon equations* IV: *The stable and unstable impulses*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[8] J. Gambaudo, P. Glendinning, and C. Tresser, *The gluing bifurcation* I: *Symbolic dynamics of the closed curves*, Nonlinearity, 1 (1988), pp. 203–214.

[9] D. Henry, *The Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[10] H. Ikeda, *Singular pulse wave bifurcations from front and back waves in bistable reaction-diffusion-systems*, Methods Appl. Anal., 3 (1996).

[11] H. Ikeda, *Existence and stability of pulse waves bifurcated from front and back waves in reaction-diffusion systems*, preprint.

[12] C. Jones, *Stability of the travelling wave solution of the Fitzhugh–Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[13] Y. Kan-on, *Parameter dependence of propagation speed travelling waves for competition-diffusion equations*, SIAM J. Math. Anal., 26 (1995), pp. 340–363.

[14] Y. Kan-on, *Existence of standing waves for competition-diffusion equations*, Japan J. Indust. Appl. Math., 13 (1996), pp. 117–133.

[15] Y. Kan-on, *Instability of stationary solutions for Lotka–Volterra competition model with diffusion*, preprint.

[16] Y. Kan-on and Q. Fang, *Stability of monotone travelling waves for competition-diffusion equations*, Japan J. Indust. Appl. Math., 3 (1996), pp. 343–349.

[17] H. Kokubu, *Homoclinic and heteroclinic bifurcation of vector fields*, Japan J. Appl. Math., 5 (1988), pp. 455–501.

[18] H. Kokubu, Y. Nishiura, and H. Oka, *Heteroclinic and homoclinic bifurcations in bistable reaction diffusion systems*, J. Differential Equations, 86 (1990), pp. 260–341.

[19] S. Nii, *N-homoclinic bifurcations for homoclinic orbits changing their twisting*, J. Dynamics Differential Equations, to appear.

[20] E. Yanagida, *Stability of fast travelling pulse solutions of the Fitzhugh–Nagumo equations*, J. Math. Biol., 22 (1985), pp. 81–104.

[21] E. Yanagida and K. Maginu, *Stability of double-pulse solutions in nerve axon equations*, SIAM J. Appl. Math., 49 (1989), pp. 1158–1173.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 434–451, March 1997

SOLUTION OF A FINITE CONVOLUTION EQUATION WITH A HANKEL KERNEL BY MATRIX FACTORIZATION*

NORBERT GORENFLO† AND MATTHIAS WERNER‡

This paper is dedicated to the memory of our friend Hans-Jürgen Böttger, who worked at the Department of Mathematics at the Technical University of Berlin. He spent a great deal of time in mathematical discussions with us.

Abstract. The following problem is considered. The convolution f of a function g supported in the interval $[-1, 1]$ with the function $H_0(k|\cdot|)$ is known on $[-1, 1]$. An expression for g is searched for. It is shown that the problem of continuing the convolution f from the interval $[-1, 1]$ to the whole real axis in a consistent way is equivalent to solving a certain Hilbert boundary-value problem for two unknown functions. This Hilbert boundary-value problem differs essentially from the corresponding one from the modern theory of finite convolution equations in Sobolev spaces (cf. [B. V. Pal'cev, *Math. USSR Sb.*, 41 (1982), pp. 289–328]), which has not yet been factored, in that it is set up in the original space, not in the range of the Fourier transform operator, and it does not contain the full convolution kernel but only its asymptotics at infinity. It is shown that a factorization for this problem can be given in terms of solutions of a certain singular algebraic ordinary differential equation. This factorization leads to an integral representation of the unknown function g. Finally, the singular differential equation, which remains to be solved, is discussed. At this point, work should be continued.

Key words. convolution equations on a finite interval, factorization of matrix functions, singular differential equations

AMS subject classifications. 45E10, 45H05, 30E20, 34B30

PII. S0036141095289154

1. Introduction.

In this paper, a factorization method for solving the finite convolution equation of the first kind,

$$\int_{-1}^{1} g(y)H_0(k|x-y|)\,dy = f(x), \quad |x| \le 1, \tag{1.1}$$

is presented, where H_0 denotes the first-kind Hankel function of order zero.

In spite of a highly developed Sobolev space theory of finite convolution equations, which has been worked out during the last 15 years and is based on the equivalence of such a convolution equation with a corresponding matrix Hilbert boundary-value problem (cf. [1] and, especially, the profound paper of Pal'cev [15]), the problem of factorization of the Hilbert boundary-value problem which is assigned to equation (1.1) according to this theory remains an open question (regarding the hitherto unsatisfactory results in the attempts of solving equation (1.1) by Wiener–Hopf methods, cf. the survey of Meister [9, p. 226]). Thus the only closed-form solution of (1.1) which was known up to the present is a "classical" solution. It consists of a series representation of g in terms of the eigenfunctions of the integral operator Q,

$$QF(\alpha) = \int_0^{\pi} F(\beta)H_0(k|\cos\alpha - \cos\beta|)\,d\beta, \quad \alpha \in [0, \pi],$$

*Received by the editors July 17, 1995; accepted for publication (in revised form) December 28, 1995.

http://www.siam.org/journals/sima/28-2/28915.html

†Siemensstraße 17, 10551 Berlin, Germany.

‡Systemintegration, VDI/VDE-Technologiezentrum Informationstechnik GmbH, Rheinstraße 10B, 14513 Teltow, Germany.

which is connected to (1.1) by the substitutions $x = \cos\alpha$, $y = \cos\beta$. The eigenfunctions of Q are the even Mathieu functions $ce_n(\cdot, k^2/4)$, $n \in N_0$; here we have used the notation in [11]. For the approach of representing the solution g of (1.1) as a series of Mathieu functions, see, e.g., [4] and [2].

In the present paper an integral representation of the solution of (1.1) in terms of the right-hand side f and solutions of a certain singular algebraic ordinary differential equation will be derived. This differential equation is of higher complexity than the standard ordinary differential equations of mathematical physics, including Mathieu's equation and the differential equation for spheroidal functions, and remains to be studied.

The solutions of the differential equation used in our approach are obtained by the factorization of a matrix Hilbert boundary-value problem for two unknown functions whose solution can be used to extend the right-hand side f of (1.1) from the interval $[-1, 1]$ to the whole real axis in a consistent way, which means that (1.1) holds for all $x \in \mathbb{R}$ (so that (1.1) can be solved by Fourier transformation).

This Hilbert problem differs essentially from the Hilbert problems occurring in the above-mentioned newer theory of finite convolution equations and, respectively, boundary-value problems for partial differential equations (cf. [1], [9], [10], [15], [16], and the references therein). Whereas the latter Hilbert problems are set up in the range of the Fourier transform operator, the Hilbert problem used in this paper is set up in the original space. Also, the coefficient matrix of the Hilbert problems considered in [1], etc., contains the symbol of the underlying convolution kernel, while the coefficient matrix of ours does not contain the full kernel $H_0(k|\cdot|)$ but is based only on its asymptotics at infinity, which causes a considerable simplification of the Hilbert problem.

We have divided the process of the solution of (1.1) in section 2 into three steps. In step 1, we establish the Hilbert problem; in step 2, we show how it can be factored by the use of solutions of an ordinary differential equation; and finally, in step 3, we give the solution of (1.1). In section 3, the ordinary differential equation on which our solution of (1.1) is founded is specified and discussed.

Our solution of (1.1) holds for arbitrary $k \neq 0$ with $\operatorname{Re} k \geq 0$ and $\operatorname{Im} k \geq 0$. On this general condition for every $p > 1$ and $f \in L^p([-1, 1])$) equation (1.1) has at most one solution: $g \in L^p([-1, 1])$. Although this special injectivity result follows from the general theory of finite convolution equations discussed above, we present a proof of it in the appendix because we need some of the formulas therein for our final solution of (1.1).

We still want to note that the factorization of the Hilbert boundary-value problem presented in this paper by solutions of an ordinary differential equation is closely related to the generalized Riemann problem examined in [3].

We also want to cite [5], where (by function-theoretic methods) finite convolution equations with simpler kernels are solved explicitly.

2. Solution of the integral equation.

To begin with, we recall that, regarding the parameter k in (1.1), we assume

$$k \neq 0, \quad \operatorname{Re} k \geq 0, \quad \text{and} \quad \operatorname{Im} k \geq 0.$$

Throughout this paper, we assume that the function g in (1.1) lies in $L^p([-1, 1])$ for some $p > 1$.

Step 1. We reformulate problem (1.1) as the problem of solving a certain Hilbert boundary-value problem for two unknown functions. With the solution of this Hilbert problem, we shall be able to express the right-hand side f in (1.1) for arguments $x \in \mathbb{R}\backslash[-1,1]$ as well.

Theorem 2.1. *If g is a solution of* (1.1), *we have*

$$\mu^-(x) + A(x)\mu^+(x) = \chi_{(-1,1)}(x) \begin{pmatrix} e^{-ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(1)}(k|x-y|)\,dy \\ e^{ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(1)}(k|x-y|)\,dy \end{pmatrix} \tag{2.1}$$

$$= \chi_{(-1,1)}(x) f(x) \begin{pmatrix} e^{-ikx} \\ e^{ikx} \end{pmatrix}, \quad x \in \mathbb{R}\backslash\{-1,1\},$$

with

$$\chi_{(-1,1)}(x) = 1 \quad \textit{for } |x| < 1, \qquad \chi_{(-1,1)}(x) = 0 \quad \textit{for } |x| > 1,$$

$$A(x) = \begin{pmatrix} 1 & -2e^{-2ikx} \\ 0 & -1 \end{pmatrix} \quad \textit{for } x < -1,$$

$$A(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{for } |x| < 1,$$

$$A(x) = \begin{pmatrix} -1 & 0 \\ -2e^{2ikx} & 1 \end{pmatrix} \quad \textit{for } x > 1,$$

$$\mu^-(x) = \frac{1}{2} \begin{pmatrix} e^{-ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(1)}(k(x-y))\,dy \\ -e^{ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(2)}(k(x-y))\,dy \end{pmatrix} \quad \textit{for } \operatorname{Im} x \le 0, \tag{2.2}$$

and

$$\mu^+(x) = \frac{1}{2} \begin{pmatrix} -e^{-ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(2)}(-k(x-y))\,dy \\ e^{ikx} \displaystyle\int_{-1}^{1} g(y) H_0^{(1)}(-k(x-y))\,dy \end{pmatrix} \quad \textit{for } \operatorname{Im} x \ge 0. \tag{2.3}$$

μ^- is holomorphic in the lower complex half-plane $\operatorname{Im} x < 0$, μ^+ is holomorphic in the upper complex half-plane $\operatorname{Im} x > 0$, and both functions are continuous up to the boundary line $\mathbb{R}$. Furthermore, we have

$$\mu^-(x) = O\left(|x|^{-\frac{1}{2}}\right) \quad \textit{and} \quad \mu^+(x) = O\left(|x|^{-\frac{1}{2}}\right), \quad |x| \to \infty. \tag{2.4}$$

Proof. The Hankel functions $H_\nu^{(1)}$ and $H_\nu^{(2)}$ of first and second kind, respectively, and order ν satisfy the following monodromic relations in $\mathbb{C}\backslash(-\infty, 0]$:

$$H_\nu^{(1)}(e^{m\pi i}z) = \frac{\sin(1-m)\nu\pi}{\sin\nu\pi} H_\nu^{(1)}(z) - e^{-\nu\pi i}\frac{\sin m\nu\pi}{\sin\nu\pi} H_\nu^{(2)}(z),$$

$$H_\nu^{(2)}(e^{m\pi i}z) = \frac{\sin(1+m)\nu\pi}{\sin\nu\pi} H_\nu^{(2)}(z) + e^{\nu\pi i}\frac{\sin m\nu\pi}{\sin\nu\pi} H_\nu^{(1)}(z),$$

$$m \in \mathbb{Z} \quad \text{(see [6, 8.476])}.$$

From these equations for $\nu = 0$ and $m = -1$, respectively, $m = 1$, we conclude that for $|y| < 1$,

$$\begin{aligned}
H_0^{(1)}(k(x-y))_- - H_0^{(2)}(-k(x-y))_+ - 2H_0^{(1)}(-k(x-y))_+ &= 0 \quad \text{for } x < -1,\\
-H_0^{(2)}(k(x-y))_- - H_0^{(1)}(-k(x-y))_+ &= 0 \quad \text{for } x < -1,\\
H_0^{(1)}(k(x-y))_- - H_0^{(2)}(-k(x-y))_+ &= 2H_0^{(1)}(k|x-y|) \quad \text{for } |x| < 1,\\
-H_0^{(2)}(k(x-y))_- + H_0^{(1)}(-k(x-y))_+ &= 2H_0^{(1)}(k|x-y|) \quad \text{for } |x| < 1,\\
H_0^{(1)}(k(x-y))_- + H_0^{(2)}(-k(x-y))_+ &= 0 \quad \text{for } x > 1,\\
-H_0^{(2)}(k(x-y))_- + 2H_0^{(2)}(-k(x-y))_+ + H_0^{(1)}(-k(x-y))_+ &= 0 \quad \text{for } x > 1.
\end{aligned}$$

Here the subscript "$-$" or "$+$" means that x approaches the real axis from the lower or, respectively, upper complex half-plane. Multiplication by $g(y)$, integration over $(-1, 1)$, and balancing with the exponential factors e^{-ikx} and e^{ikx} now yields (2.1).

From the asymptotic behavior of the Hankel functions [6, 8.451], we obtain (2.4). □

If problem (2.1) is solved, we can compute the right-hand side f in (1.1) for the values $x \in \mathbb{R}\backslash[-1, 1]$ as follows:

$$f(x) = \int_{-1}^{1} g(y) H_0(k|x-y|)\, dy = 2\begin{cases} e^{-ikx}\mu_2^+(x), & x < -1,\\ e^{ikx}\mu_1^-(x), & x > 1. \end{cases} \tag{2.5}$$

Thus f is known on the whole real axis and (1.1) can be solved by Fourier transformation.

Step 2. We derive a fundamental system for the homogeneous problem

$$\mu^-(x) + A(x)\mu^+(x) = 0, \quad x \in \mathbb{R}\backslash\{-1, 1\}. \tag{2.6}$$

To this end, the theory of Vekua for the Hilbert boundary-value problems with piecewise-continuous coefficients for several unknown functions [19, Chapter 2] is used. The piecewise-continuous coefficients are the components of the matrix A.

Throughout the following, it is assumed that $\operatorname{Im} k > 0$ because then the matrix A has definite limits for $x \to \pm\infty$. Later in this paper, we shall show that all of the results that we shall derive for the case $\operatorname{Im} k > 0$ also hold in the case $\operatorname{Im} k = 0$, i.e., $k > 0$.

The transformation of (2.6) to a Hilbert problem on a closed contour with the transformation

$$z = -\frac{ix}{x-i}$$

leads to

$$\Phi^+(z) = G(z)\Phi^-(z), \quad z \in L. \tag{2.7}$$

Here L is the boundary of the circle with radius 1/2 and center $-i/2$,

$$\Phi^+(z) = \mu^-(x), \qquad \Phi^-(z) = \mu^+(x),$$

where Φ^+ is defined inside and Φ^- is defined outside L and

$$G(z) = \begin{cases} \begin{pmatrix} 1 & 0 \\ 2e^{-\frac{2kz}{z+i}} & -1 \end{pmatrix}, & z \in a_3 \frown a_1, \\ \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, & z \in a_1 \frown a_2, \\ \begin{pmatrix} -1 & 2e^{\frac{2kz}{z+i}} \\ 0 & 1 \end{pmatrix}, & z \in a_2 \frown a_3 \end{cases}$$

with

$$a_1 = \frac{1-i}{2}, \qquad a_2 = \frac{-1-i}{2}, \quad \text{and} \quad a_3 = -i.$$

$a_p \frown a_q$ is the segment of L joining the points a_p and a_q in the mathematical positive direction. The discontinuities $z = a_1$, $z = a_2$, and $z = a_3$, respectively, of G correspond to the discontinuities $x = 1$, $x = -1$, and $x = \infty$, respectively, of A.

Problem (2.7) fulfills all the conditions in [19, Chapter 2]. With the notation therein, obviously a_1, a_2, and a_3 are all nonspecial points of (2.7) and the corresponding numbers ρ_j^σ are all equal to 0, respectively, $\pm 1/2$.

We consider the solutions Φ of (2.7) of class $h(a_1, a_2)$. This means that

$$\lim_{z\in\mathbb{C}, z\to a_1} |z-a_1|^\varepsilon \Phi(z) = 0, \qquad \lim_{z\in\mathbb{C}, z\to a_2} |z-a_2|^\varepsilon \Phi(z) = 0 \quad \text{for all } \varepsilon > 0$$

$$\text{and} \tag{2.8}$$

$$|\Phi(z)| \le c|z-a_3|^{-\alpha}, \quad z \in \mathbb{C} \text{ near } a_3, \quad \text{for some } c > 0 \text{ and } \alpha \in [0,1).$$

In the last estimate, $|\Phi(z)|$ denotes any norm of the vector $\Phi(z)$.

LEMMA 2.1. *The partial indices of the Hilbert problem* (2.7), (2.8) *are both zero.*

Proof. From [19, formula (13.49)], we obtain that the total index of (2.7) and (2.8) is zero. Thus it remains to show that both partial indices are nonnegative.

Because of (2.1), for any function f defined by a function $g \in L^p([-1,1])$ for some $p > 1$ via (1.1) with

$$\varphi^+(z) = (x(z)-i)\mu^-(x(z)), \qquad \varphi^-(z) = (x(z)-i)\mu^+(x(z))$$

and

$$b(z) = \chi_{(-1,1)}(x(z))(x(z)-i)f(x(z))(e^{-ikx(z)}, e^{ikx(z)})^T, \quad z \in L, \tag{2.9}$$

we have

$$\varphi^+(z) = G(z)\varphi^-(z) + b(z), \quad z \in L,$$

and because of (2.2), (2.3), and (2.4) the function φ belongs to the class $h(a_1, a_2)$. Furthermore, $\varphi(z)$ is vanishing for $|z| \to \infty$.

It is not hard to show that for $g \in L^p([-1,1])$, $p > 1$, the function f defined by (1.1) is Hölder continuous on $[-1,1]$. (Because of the logarithmic singularity of the function H_0, there is an analogy to the logarithmic potential in two-dimensional potential theory, whose Hölder continuity is a well-known fact.) Thus it follows (see [19, formula (14.6)]) that

$$\int_L b(z)^T [X^+(z)^T]^{-1} q(z)\, dz = 0 \tag{2.10}$$

for all vectors

$$q = (q_{-\chi_1 - 1}, q_{-\chi_2 - 1})^T,$$

where X is a fundamental matrix for (2.7) and (2.8), χ_1 and χ_2 are the partial indices of (2.7) and (2.8), and q_α is an arbitrary polynomial of degree not greater than α, $q_\alpha \equiv 0$ for $\alpha < 0$.

It remains to show that if (2.10) holds for every function b of the form (2.9), the polynomials $q_{-\chi_1-1}$ and $q_{-\chi_2-1}$ both must vanish because this implies χ_1, $\chi_2 \geq 0$. For this aim we derive a contradiction from the hypothesis $q_{-\chi_1-1} \not\equiv 0$. (The hypothesis $q_{-\chi_2-1} \not\equiv 0$ is contradicted analogously.) Hence assume $q_{-\chi_1-1} \not\equiv 0$. Because the total index $\chi_1 + \chi_2$ is zero, it follows that $q_{-\chi_2-1} \equiv 0$. Thus with

$$e_1 := (1, 0)^T \quad \text{and} \quad \varphi(z) := (e^{-ikx(z)}, e^{ikx(z)})[X^+(z)^T]^{-1} e_1,$$

(2.10) yields

$$\int_{a_1 \frown a_2} f\left(\frac{iz}{z+i}\right) \frac{q_{-\chi_1-1}(z)}{z+i} \varphi(z)\, dz = 0$$

for all functions f of the form (1.1) with $g \in L^p([-1,1])$ for arbitrary $p > 1$. Because for $p > 1$ the integral operator (1.1) is injective on $L^p([-1,1])$ (as proved in the appendix) and has the same kernel as its adjoint on $L^q([-1,1])$, $1/p + 1/q = 1$, the adjoint operator is also injective and therefore the functions f defined by (1.1) for $g \in L^p([-1,1])$ are dense in $L^p([-1,1])$ [17, corollary to Theorem 4.12]. Because of $q_{-\chi_1-1} \not\equiv 0$, it follows that φ must vanish identically on $a_1 \frown a_2$. (For the growth of $[X(z)^T]^{-1}$ near $z = a_1, a_2$, see [19, p. 97].)

With

$$\det\,(X(z))[X(z)^T]^{-1} e_1 = (-\alpha_2(z), \alpha_1(z))^T,$$

by definition of a fundamental matrix, the function $\alpha := (\alpha_1, \alpha_2)^T$ is a nonzero solution of (2.7) and (2.8) (namely, the negative of the second column of X). Furthermore, it holds that

$$e^{ikx(z)} \alpha_1^+(z) - e^{-ikx(z)} \alpha_2^+(z) = \det(X^+(z))\varphi(z) = 0, \quad z \in a_1 \frown a_2,$$

or

$$\alpha^+(z) = \alpha_1^+(z)(1, e^{2ikx(z)})^T, \quad z \in a_1 \frown a_2. \tag{2.11}$$

Because the sectionally holomorphic function α is a solution of (2.7), by virtue of (2.7), it can be continued analytically from the interior of L into the region outside of L across the segment $a_1 \frown a_2$. Note that the matrix G obviously possesses a corresponding analytic continuation; that the thus-defined function α is holomorphic in a neighborhood of $a_1 \frown a_2$ is an immediate consequence from Morera's theorem [12, p. 367]. By the analyticity of α^+ in a neighborhood of $a_1 \frown a_2$, we see by analytic continuation that (2.11) also holds inside L and therefore also on the segments $a_3 \frown a_1$ and $a_2 \frown a_3$. It follows that

$$\alpha^-(z) = G(z)^{-1}\alpha^+(z) = \alpha_1^+(z)G(z)^{-1}(1, e^{2ikx(z)})^T, \quad z \in L\backslash\{a_1, a_2, a_3\}.$$

The vector $(1, e^{2ikx(z)})^T$ is an eigenvector of the matrix $G(z)^{-1} = G(z)$ to the eigenvalue 1 on $a_3 \frown a_1 \cup a_2 \frown a_3$ and to the eigenvalue -1 on $a_1 \frown a_2$. We conclude that α_1 is a nonzero solution of class $h(a_1, a_2)$ of the simple one-dimensional Hilbert problem

$$\begin{gathered}\alpha_1^+(z) = \alpha_1^-(z) \quad \text{on } a_3 \frown a_1 \text{ and } a_2 \frown a_3\\ \text{and}\\ \alpha_1^+(z) = -\alpha_1^-(z) \quad \text{on } a_1 \frown a_2.\end{gathered} \tag{2.12}$$

Because (2.11) holds for $z \in L\backslash\{a_1, a_2, a_3\}$, we have

$$\alpha_1^+(z) = e^{\frac{2kz}{z+i}}\alpha_2^+(z), \quad z \in L\backslash\{a_1, a_2, a_3\}.$$

Because $\operatorname{Im} k > 0$, it follows that $\alpha_1^+(z)$ decays exponentially if z approaches $a_3 = -i$ along the segment $a_2 \frown a_3$. However, this is impossible for the nonzero solution α_1 of (2.12) of class $h(a_1, a_2)$ (cf. [13, formula (78, 16)]). Hence our assumption $q_{-\chi_1-1} \not\equiv 0$ leads to a contradiction. □

Next, it will be shown that the fundamental solutions of (2.7) and (2.8) fulfill a certain 2×2 system of singular algebraic differential equations. To this end, we need the following estimations for the derivatives of solutions of (2.7) and (2.8).

Lemma 2.2. *If Φ is a solution of* (2.7) *and* (2.8) *we have*

$$\begin{gathered}\lim_{z\in\mathbb{C}, z\to a_1} |z-a_1|^{1+\varepsilon}\Phi'(z) = 0, \qquad \lim_{z\in\mathbb{C}, z\to a_2} |z-a_2|^{1+\varepsilon}\Phi'(z) = 0 \quad \textit{for all } \varepsilon > 0,\\ \textit{and}\\ |\Phi'(z)| \le c|z-a_3|^{-(2+\alpha)}, \quad z \in \mathbb{C} \textit{ near } a_3,\end{gathered} \tag{2.13}$$

for α from (2.8) *and some $c > 0$.*

Proof. It is useful to consider the original problem (2.6). Therefore, let $\mu(x) = \Phi(z)$. To prove the estimate for a_1, because

$$\Phi'(z) = -\frac{1}{(z+i)^2}\mu'(x), \tag{2.14}$$

it must be shown that

$$\lim_{x_0\in\mathbb{C}, x_0\to 1} |x_0 - 1|^{1+\varepsilon}\mu'(x_0) = 0 \quad \text{for all } \varepsilon > 0. \tag{2.15}$$

For x_0 near 1, $x_0 \neq 1$, $0 < r < |x_0 - 1|$, and small $\delta > 0$, we consider the function μ on the circle $B_{r,\delta} := \{x \in \mathbb{C} |\, |x - x_0| < r + \delta\}$. If $B_{r,\delta}$ intersects the real line, we continue the sectionally holomorphic function μ analytically into the lower, respectively, upper complex half-plane ("lower" if x_0 lies in the upper half-plane; "upper" if x_0 lies in the lower half-plane) by virtue of equation (2.6). From the Cauchy integral formula we now obtain

$$\mu'(x_0) = \frac{1}{2\pi i}\int_{|x-x_0|=r} \frac{\mu(x)}{(x-x_0)^2}dx;$$

hence

$$|\mu'(x_0)| \leq \frac{\max\limits_{|x-x_0|=r} |\mu(x)|}{r}. \tag{2.16}$$

However, by (2.8),

$$\max_{|x-x_0|=r} |\mu(x)| \leq c \max_{|x-x_0|=r} |x-1|^{-\varepsilon} = c(|x_0-1|-r)^{-\varepsilon} \quad \text{for all } \varepsilon > 0.$$

Now (2.15) follows from these two inequalities if we set, e.g., $r = |x_0-1|/2$. The estimate for $\Phi'(z)$ near $z = a_2$ is obtained analogously.

Because the singularity $z = a_3 = -i$ corresponds to $x = \infty$, by virtue of (2.14), it remains to show that

$$|\mu'(x_0)| \leq c|x_0|^{\alpha} \tag{2.17}$$

if x_0 tends to infinity in the lower, respectively, upper half-plane. Without loss of generality, let $\operatorname{Im} x_0 \leq 0$. For fixed $r > 0$ and $\delta > 0$, continue the function μ analytically from the lower half-plane into the region $0 < \operatorname{Im} x < r + \delta$, $|\mathrm{Re}\ x| > 1$, across the two half-lines $(-\infty, -1)$ and $(1, \infty)$ by virtue of relation (2.6). In the same way as above, we obtain (2.16) and hence, by (2.8), also (2.17). Note that in the latter case the exponential factors $e^{\pm 2ikx}$, which occur by the process of analytic continuation, behave well for $\operatorname{Re} x \to \pm\infty$ and that the radius r is the same for all values of x_0. □

After this preparation, we can prove the following theorem. Concerning the involved theory of differential equations having singular points we refer, for example, to [8, §23] (cases $x = -1$ and $x = 1$ in what follows) and [7, Chapter XIX] (case $x = \infty$).

THEOREM 2.2. *Let X be a fundamental matrix of* (2.7) *and* (2.8)*, let V, $V(x) = X(z)$, be the corresponding fundamental matrix of* (2.6)*, and define the diagonal matrix D by*

$$D = ik \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Then the matrix-valued function Y,

$$Y(x) := V(x)^T e^{xD}, \quad e^{xD} = \begin{pmatrix} e^{ikx} & 0 \\ 0 & e^{-ikx} \end{pmatrix},$$

fulfills the first-order system of differential equations

$$(x^2-1)Y'(x) = P(x)Y(x), \tag{2.18}$$

where P is a matrix-valued function whose component functions are polynomials of degree 2 *or less.*

The points $x = -1$ and $x = 1$ are regular singular points of (2.18) *and*

$$\textit{for } x \in \{-1, 1\}, \quad \textit{the indices at } x \textit{ are } 0 \textit{ and } \frac{1}{2}. \tag{2.19}$$

Furthermore,

(2.20) (2.18) *possesses a nonzero entire solution,* *and there are solutions φ and Ψ of* (2.18) *with the asymptotics*

$$\varphi(x) \sim x^{\frac{1}{2}} e^{-ikx} c^{(1)} \quad \textit{and} \quad \Psi(x) \sim x^{\frac{1}{2}} e^{ikx} c^{(2)}, \quad |x| \to \infty, \tag{2.21}$$

such that

$$\begin{gathered} \varphi \textit{ and } \Psi \textit{ are the solutions to the index } \frac{1}{2} \\ \textit{at the points } x = -1 \textit{ and } x = 1, \textit{ respectively.} \end{gathered} \tag{2.22}$$

Here $c^{(1)}$ and $c^{(2)}$ are nonzero constant complex vectors and each of the asymptotic relations (2.21) *holds in a certain sector of the complex plane.*

Proof. By differentiating (2.6) with V instead of μ we arrive at the central relationship

$$(V^-)'(x) + DV^-(x) + A(x)[(V^+)'(x) + DV^+(x)] = 0, \quad x \in \mathbb{R}\backslash\{-1, 1\},$$

that is, $V' + DV$ is also a solution of (2.6). Hence $-(\cdot - a_3)^2 X' + DX$ is a solution of (2.7) and so is X^*,

$$X^*(z) := (z - a_1)(z - a_2)(z - a_3)^2 X'(z) - (z - a_1)(z - a_2)DX(z).$$

Because of (2.13), X^* also fulfills condition (2.8) and hence lies in the class $h(a_1, a_2)$.

Because both partial indices of problem (2.7), (2.8) are zero (Lemma 2.1), the fundamental matrix X is holomorphic in $z = \infty$ and so X' in $z = \infty$ has a zero of order at least 2. Therefore, X^* has degree 2 or less at $z = \infty$. Hence (because the partial indices of (2.7) and (2.8) are zero and X and X^* are of the same class $h(a_1, a_2)$) there exists a matrix function Q whose coefficients are polynomials of degree 2 or less in z, with

$$X^*(z) = X(z)Q(z)$$

(see [19, formula (13.47)]).

Because of

$$X^*(z) = \frac{x^2 - 1}{2(x - i)^2}[V'(x) + DV(x)],$$

with

$$R(x) := 2(x - i)^2 Q(z) = 2(x - i)^2 Q\left(-\frac{ix}{x - i}\right)$$

it follows that

$$(x^2 - 1)[V'(x) + DV(x)] = V(x)R(x).$$

The coefficients of the matrix function R are polynomials of degree 2 or less in x.

With

$$W(x) := e^{xD}V(x), \quad e^{xD} = \begin{pmatrix} e^{ikx} & 0 \\ 0 & e^{-ikx} \end{pmatrix},$$

we arrive at

$$(x^2-1)W'(x) = W(x)R(x).$$

By transposing, we now obtain (2.18), where $Y = W^T$ and $P = R^T$.

Obviously, the points $x = -1$ and $x = 1$ are regular singular points of (2.18).

With the matrix A from (2.6), we define

$$B(x) := e^{xD}A(x)e^{-xD}, \quad x \in \mathbb{R}\backslash\{-1,1\}.$$

B is constant on each of the segments $(-\infty,-1)$, $(-1,1)$, and $(1,\infty)$.

The matrix $W = e^{\cdot D}V$ fulfills

$$W^-(x) + B(x)W^+(x) = 0, \quad x \in \mathbb{R}\backslash\{-1,1\},$$

and therefore W^-, which we think is continued analytically into the upper complex half-plane through the segment $(-1,1)$ via the relation $W^-(x) = -W^+(x)$, $x\epsilon(-1,1)$, satisfies the following monodromic relation if the point $x = -1$ is encircled in the mathematical positive direction:

$$W^-(x)_{\text{new}} = \begin{pmatrix} 1 & -2 \\ 0 & -1 \end{pmatrix} W^-(x)_{\text{old}}.$$

From this it follows that the first column of the matrix solution Y^* of (2.18),

$$Y^*(x) := \left[\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} W^-(x)\right]^T,$$

returns to its original value and that the second column of Y^* changes sign if $x = -1$ is encircled. Therefore, the indices of (2.18) at $x = -1$ are of the form n and $m + 1/2$ for some n, $m \in \mathbb{Z}$. That $n = m = 0$ follows from the fact that the fundamental matrix V of (2.6) satisfies the estimates

$$\lim_{x\in\mathbb{C}, x\to -1} |x+1|^{\varepsilon} V(x) = 0 \quad \text{for all } \varepsilon > 0$$

and

$$|V(x)^{-1}| \leq c|x+1|^{-\alpha}, \quad x \in \mathbb{C} \text{ near } -1, \quad \text{for some } c > 0 \quad \text{and} \quad \alpha \in [0,1).$$

Here $|V(x)^{-1}|$ denotes any norm of the matrix $V(x)^{-1}$; the first estimate is a consequence of (2.8) and the second estimate holds by definition of a fundamental matrix of class $h(a_1, a_2)$ [19, p. 97].

In the same way, it can be proved that the indices of (2.18) at $x = 1$ are 0 and 1/2. The solution to the index 0 at $x = 1$ turns out to be proportional to the solution to the index 0 at $x = -1$; this shows (2.20).

Because of $\operatorname{Im} k > 0$, the matrix A of (2.6) satisfies

$$\lim_{x\to+\infty} A(x)^{-1} \lim_{x\to-\infty} A(x) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.23}$$

From this, by the construction of the fundamental matrix of class $h(a_1, a_2)$ performed in [19], it can be deduced that for the fundamental matrix X of (2.7) and (2.8),

$$\lim_{z \to -i} (z+i)^{\frac{1}{2}} X(z) = X_0 \quad \text{with } \det X_0 \neq 0$$

if z approaches $-i$ within the given sectors of the complex plane. (Note that there is no logarithmic singularity near $z = -i$ because the matrix (2.23) possesses a complete set of eigenvectors to the double eigenvalue -1; the matrix X_0 is constant on each sector.) It follows that for the solution Y of (2.18),

$$Y(x) = V(x)^T e^{xD} = X(z)^T e^{xD},$$

the asymptotic relation

$$Y(x) \sim x^{\frac{1}{2}} X_0^T e^{xD}, \quad |x| \to \infty,$$

holds within the corresponding sectors. Hence (2.21) is also valid. Because the second (respectively, first) column of Y changes sign if $x = -1$ (respectively, $x = 1$) is encircled (the second column of Y is just the second column of the previously examined matrix solution Y^* of (2.18)), (2.22) is also proved. □

In section 3, an explicit expression for the polynomial matrix P is given which contains only two complex parameters whose values are still unknown.

Until now, the assumption $\operatorname{Im} k > 0$ was necessary, but now we prove the following fact.

Theorem 2.3. *All of the results obtained so far also hold in the case* $\operatorname{Im} k = 0$ *(i.e.,* $k > 0$*).*

Proof. Let $k > 0$ and $0 < \alpha < \pi/2$, and let γ be any smooth curve in the complex plane so that γ contains the segment $[-1, 1]$ as well as the set $\{\lambda e^{i\alpha} | \lambda \in \mathbb{R}, |\operatorname{Re}(\lambda e^{i\alpha})| \geq 2\}$ and that $\operatorname{Im} \gamma$ is a monotonic function of $\operatorname{Re} \gamma$. We now consider the Hilbert boundary-value problem (2.6) on the curve γ instead of $\mathbb{R}$. Again using the transformation

$$z = -\frac{ix}{x-i},$$

we obtain problem (2.7) on a closed arc L^* instead of L. Because for $k > 0$ the matrix A has definite limits for $|x| \to \infty$, $x \in \gamma$, the so-modified problem (2.7) again fulfills all the conditions in [19, Chapter 2]. Now it is not hard to see that all of the results that we have obtained so far for problem (2.6) (respectively, (2.7)) hold in exactly the same way for the modified problem (2.6) (respectively, (2.7)). We want to note that by analytic continuation from relation (2.1), it follows that

$$\mu^-(x) + A(x)\mu^+(x) = 0 \quad \text{for } x \in \gamma \backslash [-1, 1]. \qquad □$$

Step 3. We express the solution g of the original problem (1.1) in terms of the fundamental system Y of (2.18)–(2.22). Regarding the parameter k, we impose no restriction; that is, we assume $k \neq 0$, $\operatorname{Re} k \geq 0$, and $\operatorname{Im} k \geq 0$.

Theorem 2.4. *Let X be a fundamental matrix of* (2.7) *and* (2.8)*, set $V(x) = X(z)$, and define the matrix D as in Theorem* 2.2. *Consider the fundamental system Y of* (2.18)–(2.22) *specified by $Y(x) = V(x)^T e^{xD}$ for* $\operatorname{Im} x \leq 0$ *and the requirement that for* $\operatorname{Im} x \geq 0$ *the values $Y(x)$ are obtained by analytic continuation from the lower complex half-plane across the segment* $(-1, 1)$.

If f is given by (1.1), *we have*

$$f(s) = \int_{-1}^{1} f(\tau)T(s,\tau)\,d\tau \quad \textit{for all } s \in \mathbb{R}\backslash\{-1,1\}, \tag{2.24}$$

where the "transfer function" T is defined by

$$T(s,\tau) = \begin{cases} -\dfrac{1}{i\pi(s-\tau)}[Y^+(s)^T[Y(\tau)^T]^{-1}(1,1)^T]_2, & s \in (-\infty,-1), \\ \delta(s-\tau), & s \in (-1,1), \\ \dfrac{1}{i\pi(s-\tau)}[Y^-(s)^T[Y(\tau)^T]^{-1}(1,1)^T]_1, & s \in (1,\infty). \end{cases}$$

Here $[\ ]_1$ and $[\ ]_2$ denote the first and second components, respectively, of the vector within the brackets, δ is the Dirac delta distribution, and $Y^{\pm}(s)$ denotes the limit of $Y(s^*)$ if s^* approaches s from the upper or lower half-plane, respectively.

Furthermore, with

$$f\hat{}(\rho) = \int_{\mathbb{R}} f(x)e^{-i\rho x}dx,$$

the solution g of (1.1) *is given by*

$$g(x) = \frac{1}{4\pi}\int_{\mathbb{R}} \sqrt{k^2-\rho^2}\, f\hat{}(\rho)e^{i\rho x}d\rho. \tag{2.25}$$

Here the sign of the square root is specified by (A1) *in the appendix. In general, these Fourier integrals must be interpreted in the sense of tempered distributions.*

Proof. With φ^+ and φ^- defined in the formulas preceding equation (2.9) and b defined as in (2.9), it follows from (2.1) that

$$\varphi^+(z) = G(z)\varphi^-(z) + b(z)$$

for $z \in L$ (if $\operatorname{Im} k > 0$), respectively, $z \in L^*$ (if $k > 0$). Because both partial indices of (2.7) and (2.8) are zero, we have

$$\varphi(z) = \frac{1}{2\pi i}X(z)\int_{a_1\frown a_2} \frac{1}{t-z}X^+(t)^{-1}b(t)\,dt, \quad z \in \mathbb{C}$$

(see [19, formula (14.3)]). It follows that the solution μ of (2.1) is given by

$$\mu(x) = \frac{1}{2\pi i}V(x)\int_{-1}^{1}\frac{f(\tau)}{x-\tau}V^-(\tau)^{-1}(e^{-ik\tau}, e^{ik\tau})^T d\tau, \quad x \in \mathbb{C}.$$

This can be written as

$$\mu^{\mp}(x) = \pm\frac{1}{2\pi i}e^{-xD}Y(x)^T\int_{-1}^{1}\frac{f(\tau)}{x-\tau}[Y(\tau)^T]^{-1}(1,1)^T d\tau, \quad x \in \mathbb{C}.$$

(Note that for $x \in (-1,1)$, it holds that $V^+(x) = -V^-(x)$ but $Y^+(x) = Y^-(x)$ because Y has been defined by analytic continuation across $(-1,1)$.)

Together with relation (2.5), we can now express the values $f(x)$ of the right-hand side f of (1.1) for all $x \in \mathbb{R}$ as asserted in (2.24).

Formula (2.25) follows immediately from the results obtained in the appendix. □

Remark 2.1. Using (2.24), equation (2.25) can be formally written as

$$g(x) = \int_{-1}^{1} f(\tau) K(x,\tau)\, d\tau,$$

where the formal distribution kernel K is given by

$$K(x,\tau) = \frac{1}{4\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \sqrt{k^2 - \rho^2} e^{i(x-s)\rho}\, d\rho\, T(s,\tau)\, ds.$$

3. The singular differential equation. In the following theorem, we give an explicit representation of the coefficient matrix P of problem (2.18)–(2.22).

THEOREM 3.1. *There is a fundamental matrix V of* (2.6) *and* (2.8) *for which the corresponding polynomial matrix P in* (2.18) *is of the form*

(3.1)
$$P(x) = \begin{pmatrix} 0 & r - ik \\ s - ik & 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} - \sqrt{\frac{1}{4} - rs} & 0 \\ 0 & \frac{1}{2} + \sqrt{\frac{1}{4} - rs} \end{pmatrix} x + \begin{pmatrix} 0 & ik \\ ik & 0 \end{pmatrix} x^2$$

for some r, $s \in \mathbb{C}$.

Proof. We have the identity

$$ZA(-x)Z = A(x) = A(x)^{-1}, \quad x \in \mathbb{R}, \quad \text{with } Z := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Therefore, if V is any fundamental matrix of (2.6) and (2.8), the matrix function V^*,

$$(V^*)^{\mp}(x) := ZV^{\pm}(-x),$$

is also a fundamental matrix of (2.6) and (2.8). Because both partial indices of (2.6) and (2.8) are zero, there must be a constant matrix C with

$$ZV(-x) = V^*(x) = V(x)C.$$

From this, for the fundamental system Y of (2.18), which satisfies

$$Y(x) = V(x)^T e^{xD} \quad \text{for } \operatorname{Im} x \leq 0,$$

we derive

$$Y(x) = -C^T Y(-x) Z, \quad x \in (-1,1).$$

(Note that $V^+(x) = -V^-(x)$ for $x \in (-1,1)$) and $C = C^{-1}$ (by replacing x by $-x$ in the above equality and solving the resulting equation for $Y(x)$.) From (2.18), it now follows that

$$P(-x) = -C^T P(x) C^T. \tag{3.2}$$

We define the constant matrices P_0, P_1, and P_2 by

$$P(x) = P_0 + P_1 x + P_2 x^2.$$

Because of (2.19), we have the following:

$$\begin{gathered} P_0 - P_1 + P_2 \text{ has the eigenvalues } 0 \text{ and } -1 \\ \text{and} \\ P_0 + P_1 + P_2 \text{ has the eigenvalues } 0 \text{ and } 1. \end{gathered} \tag{3.3}$$

Furthermore, because of (2.21),

$$P_2 \text{ has the eigenvalues } -ik \text{ and } ik. \tag{3.4}$$

(From a more detailed analysis along the lines of the proof of the asymptotic relation (2.21) and the proof of the last estimate in (2.13), one can obtain an asymptotic relation not only for Y but also for Y', and so, using (2.18), one can directly deduce that $P_2 = X_0^T D (X_0^T)^{-1}$ with X_0 from the proof of (2.21). Note that the diagonal matrix D just has the entries $\pm ik$.)

If the fundamental matrix V is changed into the fundamental matrix VS, where S is any constant invertible matrix, the matrix C^T changes into $S^T C^T (S^T)^{-1}$. For that reason, because of $C = C^{-1}$, we can assume that C^T is equal to

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

If C^T were the first or second of these matrices, then it would follow from (3.2) that $P_2 = -P_2$ and hence that $P_2 = 0$, which would contradict (3.4). Therefore, we can assume that C^T is equal to the third of these matrices. With this form of C^T, we obtain from (3.2) that there exist $r, s, a, b, p, q \in \mathbb{C}$ with

$$P_0 = \begin{pmatrix} 0 & r - ik \\ s - ik & 0 \end{pmatrix}, \qquad P_1 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & p \\ q & 0 \end{pmatrix}. \tag{3.5}$$

Because of (3.4), we have $pq = -k^2$ so that for

$$U := \begin{pmatrix} 1 & 0 \\ 0 & \dfrac{p}{ik} \end{pmatrix},$$

it holds that

$$UP_2U^{-1} = \begin{pmatrix} 0 & ik \\ ik & 0 \end{pmatrix}.$$

Therefore, if we change the fundamental matrix V of (2.6) and (2.8), which led us to (3.5), into the fundamental matrix VU^T, the matrix P given by (3.5) changes into a matrix that also has the form (3.5) but, in addition, with $p = q = ik$.

With $p = q = ik$ in (3.5), it now follows from (3.3) that

$$a = \frac{1}{2} - \sqrt{\frac{1}{4} - rs} \quad \text{and} \quad b = \frac{1}{2} + \sqrt{\frac{1}{4} - rs},$$

and we have reached the form (3.1).

The last condition from (2.19) and (2.21) which must be fulfilled by the matrix P given by (3.1) is that the exponent in the power of x occurring in (2.21) before $e^{\mp ikx}$ is equal to $1/2$. However, this condition turns out to be fulfilled automatically if P is of the form (3.1), where r and s are arbitrary complex numbers. □

Remark 3.1. A change of the sign of the square root $(1/4 - rs)^{1/2}$ together with an exchange of r and s in the matrix P_0 simply corresponds to the change from the fundamental matrix V to the fundamental matrix

$$V \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

so that also a definite sign can be attached to the square root.

Remark 3.2. To construct the fundamental system $Y = V^T e^{\cdot D}$ of (2.18), the parameters r and s in (3.1) have to be determined so that the conditions (2.20) and (2.22) are fulfilled. Note that (2.22) leads to only one condition because if $\Psi = (\Psi_1, \Psi_2)^T$ is a solution of (2.18) to the index $1/2$ at $x = 1$, which fulfills (2.21), the function φ,

$$\varphi(x) = (\Psi_1(-x), -\Psi_2(-x))^T,$$

is the solution to the index $1/2$ at $x = -1$ and also fulfills (2.21). Because the columns of the fundamental system $Y = V^T e^{\cdot D}$ are proportional to Ψ, respectively, φ (as was shown in the proof of Theorem 2.2), for the solution of (1.1) it suffices to know the function Ψ. If for every value of k the parameters $r(k)$ and $s(k)$ are determined in the above manner, the function Ψ is a function of the parameters k and x:

$$\Psi = \Psi(k; x). \tag{3.6}$$

Finally, we want to derive a scalar differential equation of the second order which is equivalent to the first-order system (2.18).

COROLLARY 3.1. *If the vector y is a solution of* (2.18) *with the matrix P given by* (3.1)*, the first component y_1 of y satisfies the scalar second-order differential equation*

$$(x^2 - a^2)(x^2 - 1)y_1''(x) + p(x)y_1'(x) + q(x)y_1(x) = 0, \tag{3.7}$$

where

$$a = \sqrt{\frac{ir}{k} + 1},$$

$$p(x) = -x^3 + (2 - a^2)x,$$

and

$$q(x) = k^2x^4 + \left[ik(2ik - s)a^2 + \frac{1}{2} - \sqrt{\frac{1}{4} + ik(a^2 - 1)s}\right]x^2$$

$$+ ik(s - ik)a^4 + a^2\left(\frac{1}{2} - \sqrt{\frac{1}{4} + ik(a^2 - 1)s}\right).$$

Proof. With P given by (3.1) we solve the first of the two equations in (2.18) for the second component y_2 of y and differentiate the resulting expression to obtain an expression for y_2'. Then we compare the latter expression with the second equation in (2.18), in which y_2 has been replaced by the expression obtained earlier from the first equation in (2.18). This yields (3.7) with the asserted forms of a, p, and q. □

Remark 3.3. For $a^2 \neq 1$, of course, $-a$ and a are only apparent singular points of (3.7).

Remark 3.4. In the special case $a^2 = 1$ (i.e., $r = 0$) and

$$\sqrt{\frac{1}{4} + ik(a^2-1)s} = +\frac{1}{2},$$

(3.7) reduces to

$$(x^2-1)y_1''(x) - xy_1'(x) + (k^2x^2 - k^2 - iks)y_1(x) = 0.$$

This is the spheroidal differential equation with the indices 0 and 3/2 at $x = \pm 1$ [11], [12]. (The indices are not 0 and 1/2, which does not contradict (3.3) because the differential equation is only a differential equation for the first component y_1 of y.) Therefore, (3.7) is a generalization of this spheroidal differential equation. The parameters a and s, of course, have yet to be determined properly.

Remark 3.5. With respect to the importance of the functions $\Psi(k;\cdot)$, referred to in (3.6), for the solution of the original problem (1.1) by (2.24) and (2.25), we propose to thoroughly investigate the differential equation (2.18), (3.1), respectively, the differential equation (3.7).

4. Remarks. Our mathematical interest and examination of equation (1.1) found its origin in the employment of the second author with a doping problem in solid-state physics (diffusion through a slit).

For the connection between (1.1) and a certain mixed boundary-value problem for the Helmholtz equation, see, e.g., [2]. Because the diffusion equation can be transformed into a family of Helmholtz equations by performing a Laplace transform with respect to the time variable, the solution of (1.1) also leads to the solution of the corresponding mixed boundary-value problem for the diffusion equation.

Another application of equation (1.1), for example, arises in the problem of computing the distortion of a two-dimensional plane wave diffracted by a slit [18, p. 284].

In the future, we plan to demonstrate that the theory presented in this paper also applies to some other physical problems involving the Helmholtz equation and to derive the corresponding ordinary differential equations.

Appendix. We show that for $p > 1$ the integral operator

$$M : L^p([-1,1]) \to L^p([-1,1]), \qquad Mg(x) = \int_{-1}^{1} g(y)H_0(k|x-y|)\,dy,$$

is injective.

In the following, we define the Fourier transform $h\hat{}$ of a function h by

$$h\hat{}(\rho) = \int_{\mathbb{R}} h(x)e^{-i\rho x}dx, \quad \rho \in \mathbb{R}.$$

From the Sommerfeld–Weyl integral

$$H_0(k\sqrt{x^2+y^2}) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{e^{i(\sqrt{k^2-\rho^2}x+\rho y)}}{\sqrt{k^2-\rho^2}}\,d\rho, \quad x > 0,$$

with

$$\text{(A1)} \qquad \operatorname{Im} \sqrt{k^2 - \rho^2} > 0 \quad \text{if } \operatorname{Im} k > 0,$$

respectively,

$$\sqrt{k^2 - \rho^2} > 0 \quad \text{for } |\rho| < k \quad \text{and} \quad \operatorname{Im} \sqrt{k^2 - \rho^2} > 0 \quad \text{for } |\rho| > k, \quad \text{if } \operatorname{Im} k = 0$$

(see, e.g., [12, p. 823]), by letting $x \downarrow 0$, we see that (in the sense of tempered distributions)

$$\text{(A2)} \qquad H_0(k|\cdot|)\hat{\ }(\rho) = \frac{2}{\sqrt{k^2 - \rho^2}}, \quad \rho \in \mathbb{R}.$$

Now we assume that $g \in L^p([-1,1])$ for some $p > 1$ and $Mg \equiv 0$. We have to prove that $g \equiv 0$. To this end, we assume that the functions g and $f := Mg$ are defined on the whole real line by setting $g(x) = 0$ for $|x| > 1$ and defining f by (1.1) also for $|x| > 1$. Furthermore, because

$$L^r([-1,1]) \subset L^s([-1,1]) \quad \text{for } s < r,$$

we can assume that $p < 2$. (In the following, we shall make use of the theory of the Fourier transform on $L^p(\mathbb{R})$, $p < 2$, as it is given, e.g., in [14].)

Because of (A2), the Parseval formula [14, Theorem (6.4.2)] yields

$$f(x) = \int_{\mathbb{R}} g(y) H_0(k|x-y|)\, dy = \int_{\mathbb{R}} \frac{g\hat{\ }(\rho)}{\pi\sqrt{k^2 - \rho^2}} e^{i\rho x} d\rho, \quad x \in \mathbb{R}.$$

Because $g\hat{\ }$ is continuous and bounded on $\mathbb{R}$, the function $g\hat{\ }/(k^2 - \cdot^2)^{1/2}$ lies in $L^p(\mathbb{R})$. Hence repeated use of the Parseval formula leads to

$$\int_{\mathbb{R}} \frac{|g\hat{\ }(\rho)|^2}{\sqrt{k^2 - \rho^2}}\, d\rho = \int_{\mathbb{R}} \frac{g\hat{\ }(\rho)}{\sqrt{k^2 - \rho^2}} \overline{g\hat{\ }(\rho)}\, d\rho = \pi \int_{\mathbb{R}} f(x) \overline{g(x)}\, dx = 0.$$

The last equality holds because $f(x) = 0$ for $|x| \le 1$ and $g(x) = 0$ for $|x| > 1$. Together with (A1), it now follows that $g\hat{\ } \equiv 0$ and hence $g \equiv 0$.

Acknowledgment. The first author wants to thank his father, Professor Rudolf Gorenflo of the Department of Mathematics at the Free University of Berlin, for mathematical discussions and especially for making some important literature available.

REFERENCES

[1] M. A. Bastos and A. F. Dos Santos, *Convolution equations of the first kind on a finite interval in Sobolev spaces*, Integral Equations Operator Theory, 13 (1990), pp. 638–659.

[2] J. A. Belward, *The solution of an integral equation of the first kind on a finite interval*, Quart. Appl. Math., 27 (1969), pp. 313–321.

[3] G. D. Birkhoff, *The generalized Riemann problem for linear differential equations and the allied problems for linear difference and q-difference equations*, Proc. Amer. Acad. Arts Sci., 49 (1913), pp. 521–568.

[4] J. Dörr, *Zwei Integralgleichungen erster Art, die sich mit Hilfe Mathieuscher Funktionen lösen lassen*, Z. Angew. Math. Phys., III (1952), pp. 427–439.

[5] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, UK, 1966.

[6] I. S. Gradstein and I. M. Ryshik, *Tables of Series, Products and Integrals*, Vol. 2, Verlag Harri Deutsch, Thun, Frankfurt, Germany, 1981.

[7] E. L. Ince, *Ordinary Differential Equations*, Dover, New York, 1956.

[8] E. Kamke, *Differentialgleichungen* I, in Gewöhnliche Differentialgleichungen, Akademische Verlagsgesellschaft, Geest und Portig K.-G., Leipzig, Germany, 1962.

[9] E. Meister, *Einige gelöste und ungelöste kanonische Probleme der mathematischen Beugungstheorie*, Exposition. Math., 5 (1987), pp. 193–237.

[10] E. Meister, F. Penzel, F.-O. Speck, and F. S. Teixeira, *Two-media scattering problems in a half-space*, Preprint 1368, Fachbereich Mathematik, Technische Hochschule Darmstadt, Darmstadt, Germany, 1991.

[11] J. Meixner and F. W. Schäfke, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer-Verlag, Berlin, 1954.

[12] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part I, McGraw–Hill, New York, Toronto, London, 1953.

[13] N. I. Muschelischwili, *Singuläre Integralgleichungen*, Akademie-Verlag, Berlin, 1965.

[14] G. O. Okikiolu, *Aspects of the Theory of Bounded Integral Operators in L^p-Spaces*, Academic Press, London, New York, 1971.

[15] B. V. Pal'cev, *A generalization of the Wiener–Hopf method for convolution equations on a finite interval with symbols having power-like asymptotics at infinity*, Math. USSR-Sb., 41 (1982), pp. 289–328.

[16] F. Penzel, *Sobolev space methods for dual integral equations in axialsymmetric screen problems*, SIAM J. Math. Anal., 23 (1992), pp. 1167–1181.

[17] W. Rudin, *Functional Analysis*, Tata/McGraw–Hill, New Delhi, India, reprinted 1979.

[18] A. Sommerfeld, *Vorlesungen über Theoretische Physik*, Band 4, Optik, Dieterich'sche Verlagsbuchhandlung, Wiesbaden, Germany, 1950.

[19] N. P. Vekua, *Systems of Singular Integral Equations*, P. Noordhoff, Groningen, The Netherlands, 1967.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 452–480, March 1997

MULTIVARIABLE BIG AND LITTLE q-JACOBI POLYNOMIALS[*]

JASPER V. STOKMAN[†]

Abstract. A four-parameter family of multivariable big q-Jacobi polynomials and a three-parameter family of multivariable little q-Jacobi polynomials are introduced. For both families, full orthogonality is proved with the help of a second-order q-difference operator which is diagonalized by the multivariable polynomials. A link is made between the orthogonality measures and R. Askey's q-extensions of Selberg's multidimensional beta-integrals.

Key words. big q-Jacobi polynomials, little q-Jacobi polynomials, BC_n-type Askey–Wilson polynomials, multivariable orthogonal polynomials, q-extensions of Selberg's multidimensional beta-integrals

AMS subject classifications. 33D45, 33D80

PII. S0036141095287192

1. Introduction.

In the one-variable case, big (resp. little) q-Jacobi polynomials depend apart from q on (essentially) three (resp. two) parameters. The big and little q-Jacobi polynomials are orthogonal with respect to inner products which are both given by a Jackson (q-) integral over a positive weight function. The associated orthogonality measures therefore have positive weights on infinitely many discrete mass points.

The one-variable big and little q-Jacobi polynomials are q-analogues of the classical Jacobi polynomials in the sense that when q tends to 1, the big and little q-Jacobi polynomials tend (up to a possible translation and dilation of the variable) to the classical Jacobi polynomials.

The families of one-variable big and little q-Jacobi polynomials are members of the Askey–Wilson hierarchy. The Askey–Wilson hierarchy consists of families of orthogonal polynomials which are joint eigenfunctions of a second-order q-difference operator. Some families can be obtained from others by limit transitions or by specializations of parameters. This induces the hierarchy structure between the families. In this point of view, the four-parameter family of Askey–Wilson polynomials is on top of the hierarchy and the families of big (resp. little) q-Jacobi polynomials are directly below the Askey–Wilson polynomials. Suitable limit transitions are known from the Askey–Wilson polynomials to the big (resp. little) q-Jacobi polynomials (cf. [12]). Furthermore, the little q-Jacobi polynomials can be obtained from the big q-Jacobi polynomials by a suitable limit transition.

Recently, Koornwinder introduced in [11] a multivariable (BC_n-type) generalization of the family of Askey–Wilson polynomials by extending the three-parameter family of Macdonald polynomials of type (BC_n, B_n) to a five-parameter family of orthogonal polynomials. Four of these parameters play the same role as in the one-variable case, while the fifth parameter is an extra deformation parameter. The BC_n-type Askey–Wilson polynomials are again joint eigenfunctions of a second q-difference operator. Koornwinder remarked in [11] that the whole Askey–Wilson hierarchy could probably be generalized to the BC_n case as well as the limit transitions between the families.

[*] Received by the editors June 5, 1995; accepted for publication November 29, 1995.
http://www.siam.org/journals/sima/28-2/28719.html

[†] Faculty WINS, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (jasper@fwi.uva.nl).

In this paper, a four-parameter family of multivariable BC_n-type big q-Jacobi polynomials and a three-parameter family of multivariable BC_n-type little q-Jacobi polynomials are introduced which have the following properties:

(1) compared with the one-variable case, there is an extra deformation parameter involved;

(2) the multivariable big (resp. little) q-Jacobi polynomials are joint eigenfunctions of a second-order q-difference operator;

(3) the multivariable big (resp. little) q-Jacobi polynomials are mutually orthogonal with respect to an inner product, which is essentially given by a multidimensional Jackson integral over a positive weight function;

(4) in a paper by the author and Koornwinder (cf. [18]), limit transitions from multivariable Askey–Wilson polynomials to multivariable big (resp. little) q-Jacobi polynomials and from multivariable big q-Jacobi polynomials to multivariable little q-Jacobi polynomials are proved which generalize the limit transitions in the one-variable case, and it is proved that the multivariable big (resp. little) q-Jacobi polynomials are q-analogues of generalized Jacobi polynomials (see [19]) (which are related with BC_n-type Heckman–Opdam polynomials by a suitable change of variables).

This paper is organized as follows. In section 2, the definitions of big and little q-Jacobi polynomials in one variable are given. In section 3, we will consider two formal limits of the BC_n-type Askey–Wilson polynomials, which generalize the limits from Askey–Wilson polynomials to big (resp. little) q-Jacobi polynomials in the one-variable case. We will obtain two second-order q-difference operators D_B (resp. D_L), and in section 4, it will be proved that D_B and D_L are triangular with respect to the basis of monomial symmetric functions. In section 5, the multivariable big and little q-Jacobi polynomials will be introduced. We will use techniques introduced by Macdonald in [15] to prove full orthogonality of the polynomials. First, it will be proved that the big (resp. little) q-Jacobi polynomials are joint eigenfunctions of D_B (resp. D_L) by proving the self-adjointness of D_B (resp. D_L). Full orthogonality of the big (resp. little) q-Jacobi polynomials will then be a consequence of the fact that the eigenvalues are sufficiently different.

Furthermore, it will be shown in section 5 that for special values of the extra deformation parameter, the multidimensional Jackson integrals over the weight functions are essentially the q-extensions of Selberg's multidimensional beta-integrals which were introduced by Askey [3]. Askey's conjectured evaluations of these multidimensional Jackson integrals have recently been proved [6], [8], [10]. Section 6 contains some proofs which were omitted in section 5.

Notation and conventions. Throughout this paper, we work with a fixed $q \in (0,1)$. $\mathbb{N} = \{1, 2, \ldots\}$ denotes the natural numbers and $\mathbb{N}_0$ denotes the natural numbers together with 0. The convention will be used that $\prod_{i=l}^{k} a_i = 1$ if $k < l$ for $k, l \in \mathbb{N}_0$. If there is no confusion possible, the dependence on the parameters will be omitted in the formulas.

2. One variable big and little q-Jacobi polynomials.

Let $a, b \in \mathbb{R}$, $a < b$, and f be a function defined on the points $\{aq^k, bq^k \mid k \in \mathbb{N}_0\}$. Define the Jackson ($q$-) integral of f over $[a, b]$ by

$$\int_a^b f(x)\, d_qx := \int_0^b f(x)\, d_qx - \int_0^a f(x)\, d_qx,$$
$$\int_0^b f(x)\, d_qx := (1-q) \sum_{k=0}^{\infty} f(bq^k) bq^k,$$

provided that the infinite sums in the definition of the q-integral from 0 to a and in the definition of the q-integral from 0 to b are absolutely convergent. In the special case that $a = bq^{k+1}$ for some $k \in \mathbb{N}_0$, we have

$$\int_{bq^{k+1}}^{b} f(x)\, d_qx = (1-q) \sum_{m=0}^{k} f(bq^m) bq^m, \tag{2.1}$$

so we can then use (2.1) as definition of the q-integral from bq^{k+1} to b without worrying about convergence.

Define the q-shifted factorial by

$$(a;q)_b := \frac{(a;q)_\infty}{(q^b a;q)_\infty}, \qquad (a;q)_\infty := \prod_{k=0}^{\infty}(1-aq^k),$$

for $a \in \mathbb{C}$ and $b \in \mathbb{C}\backslash\mathbb{N}_0$ such that $q^b a \neq q^{-k}$ for all $k \in \mathbb{N}_0$. For $l \in \mathbb{N}_0$, we set $(a;q)_l := \prod_{k=0}^{l-1}\left(1-aq^k\right)$. Denote

$$(a_1, \ldots, a_r; q)_b := \prod_{j=1}^{r} (a_j;q)_b\,.$$

Let $c, d > 0$, $a \in (-c/dq, 1/q)$, and $b \in (-d/cq, 1/q)$ or $a = cz$ and $b = -d\bar{z}$ with $z \in \mathbb{C} \setminus \mathbb{R}$. Denote V_B^q for the set of parameters (a,b,c,d) which satisfy these conditions. Define

$$w_B(x;a,b,c,d;q) := \frac{(qx/c, -qx/d;q)_\infty}{(qax/c, -qbx/d;q)_\infty}; \tag{2.2}$$

then $w_B(x;a,b,c,d;q)$ is positive for $x \in [-d,c]$, and

$$\langle f,g\rangle_{B,1,q}^{a,b,c,d} := \int_{-d}^{c} f(x)g(x)w_B(x;a,b,c,d;q)\, d_qx, \quad f,g \in \mathbb{R}[x], \tag{2.3}$$

is a well-defined inner product on $\mathbb{R}[x]$.

Definition 2.1. *The big q-Jacobi polynomials* $\{P_m(x;a,b,c,d;q)\,|\, m \in \mathbb{N}_0\}$ *are defined by the following two conditions:*

(1) $P_m(x)$ *is a monic polynomial of degree* m *in* x*;*

(2) $\langle P_m(x), x^l\rangle_{B,1} = 0$ *if* $l < m$.

Consequently, the big q-Jacobi polynomials are mutually orthogonal with respect to $\langle\,.\,,\,.\,\rangle_{B,1}$. Explicit expressions for the big q-Jacobi polynomials are given by

$$P_m(x;a,b,c,d;q) = \frac{(qa;q)_m\,(-qad/c;q)_m}{(q^{m+1}ab;q)_m\,(qa/c)^m}\,{}_3\phi_2\left[\begin{matrix} q^{-m}, q^{m+1}ab, qxa/c \\ qa, -qad/c\end{matrix}; q, q\right],$$

with the q-hypergeometric series defined by

$${}_{r+1}\phi_r\left[\begin{matrix} a_1, \ldots, a_{r+1} \\ b_1, \ldots, b_r\end{matrix}; q, z\right] := \sum_{k=0}^{\infty} \frac{(a_1, \ldots, a_{r+1};q)_k\, z^k}{(b_1, \ldots, b_r, q;q)_k}$$

(cf. [2]).

Note that $P_m(x;a,b,c/d,1;q) = d^{-m}P_m(dx;a,b,c,d;q)$, so the big q-Jacobi polynomials depend (apart from q) essentially on a, b, and the ratio c/d. The second-order q-difference operator

$$\begin{aligned}(2.4)\quad \left(D_{1,q}^{a,b,c,d}f\right)(x) := q\left(a-\frac{c}{qx}\right)\left(b+\frac{d}{qx}\right)(f(qx)-f(x)) \\ +\left(1-\frac{c}{x}\right)\left(1+\frac{d}{x}\right)\left(f(q^{-1}x)-f(x)\right) \quad (f\in\mathbb{R}[x])\end{aligned}$$

is diagonalized by the big q-Jacobi polynomials

$$(2.5)\qquad \left(D_{1,q}^{a,b,c,d}P_m(\,.\,;a,b,c,d;q)\right)(x) = a_m^{a,b,q}P_m(x;a,b,c,d;q) \quad \forall m\in\mathbb{N}_0$$

with eigenvalues

$$(2.6)\qquad a_m^{a,b,q} := qab(q^m-1)+(q^{-m}-1).$$

Note that D_1 is self-adjoint with respect to $\langle\,.\,,\,.\,\rangle_{B,1}$ because $\{P_m(x)\,|\,m\in\mathbb{N}_0\}$ is an orthogonal basis of $\mathbb{R}[x]$ with respect to $\langle\,.\,,\,.\,\rangle_{B,1}$ which consists of eigenfunctions of D_1.

The little q-Jacobi polynomials can be introduced in a similar way. Let $0<a<1/q$ and $b<1/q$, and denote by V_L^q the set of parameters (a,b) which satisfy these conditions. Define

$$(2.7)\qquad v_L(x;a,b;q) := \frac{(qx;q)_\infty}{(qbx;q)_\infty}x^\alpha \quad (a=q^\alpha);$$

then $v_L(x;a,b;q)$ is positive for $x\in[0,1]$ and

$$(2.8)\qquad \langle f,g\rangle_{L,1,q}^{a,b} := \int_0^1 f(x)g(x)v_L(x;a,b;q)\,d_qx, \quad f,g\in\mathbb{R}[x],$$

is an inner product on $\mathbb{R}[x]$.

Definition 2.2. *The little q-Jacobi polynomials $\{p_m(x;a,b;q)\,|\,m\in\mathbb{N}_0\}$ are defined by the following two conditions:*

(1) $p_m(x)\in\mathbb{R}[x]$ *is a monic polynomial of degree m in x;*

(2) $\langle p_m(x),x^l\rangle_{L,1}=0$ *if $l<m$.*

Consequently, the little q-Jacobi polynomials are mutually orthogonal with respect to $\langle\,.\,,\,.\,\rangle_{L,1}$. Explicit expressions for the little q-Jacobi polynomials are given by

$$p_m(x;a,b;q) := \frac{(-1)^m q^{\binom{m}{2}}(qa;q)_m}{(q^{m+1}ab;q)_m}\,{}_2\phi_1\left[\begin{matrix}q^{-m},q^{m+1}ab\\ qa\end{matrix};q,qx\right]$$

(cf. [1]). The little q-Jacobi polynomials are eigenfunctions of the q-difference operator $D_{1,q}^{b,a,1,0}$ with the same eigenvalues as in the big q-Jacobi case ((2.5) and (2.6)):

$$(2.9)\qquad \left(D_{1,q}^{b,a,1,0}p_m(\,.\,;a,b;q)\right)(x) = a_m^{a,b,q}p_m(x;a,b;q) \quad \forall m\in\mathbb{N}_0,$$

so $D_{1,q}^{b,a,1,0}$ is self-adjoint with respect to $\langle\,.\,,\,.\,\rangle_{L,1,q}^{a,b}$.

Remark 2.3. In this section, we defined the one-variable big and little q-Jacobi polynomials as monic polynomials because the multivariable generalizations will be monic. However, it is more common to normalize the big and little q-Jacobi polynomials differently. The big q-Jacobi polynomials are usually defined by

$$\tilde{P}_m(x;a,b,c,d;q) = {}_3\phi_2\left[\begin{matrix} q^{-m}, q^{m+1}ab, qxa/c \\ qa, -qad/c \end{matrix}; q, q\right]$$

and the little q-Jacobi polynomials are usually defined by

$$\tilde{p}_m(x;a,b;q) = {}_2\phi_1\left[\begin{matrix} q^{-m}, q^{m+1}ab \\ qa \end{matrix}; q, qx\right].$$

For more details about the one-variable big and little q-Jacobi polynomials, see [1], [2], [7], and [13].

3. Formal limits of multivariable Askey–Wilson polynomials. Let A be the algebra of Laurent polynomials in the independent indeterminates $x_1, \ldots, x_n$. The Weyl group W corresponding to the root system of type BC_n acts in a natural way on A. Let A^W be the subalgebra of A consisting of W-invariant Laurent polynomials. Let P^+ be the partitions of length $\leq n$, so

$$P^+ := \{\lambda = (\lambda_1, \ldots, \lambda_n) \,|\, \lambda_1 \geq \cdots \geq \lambda_n \geq 0\}. \tag{3.1}$$

W acts on $\mathbb{Z}^n$ by sign changes and permutations of the coordinates. The monomials $\{\tilde{m}_\lambda \,|\, \lambda \in P^+\}$, with $\tilde{m}_\lambda := \sum_{\mu \in W\lambda} x^\mu$, form a basis of A^W. Let $a, b, c, d, t \in \mathbb{C}$ and define the weight function $\delta(x;a,b,c,d;q,t)$ by

$$\delta(x_1, \ldots, x_n) := \delta^+(x_1, \ldots, x_n)\delta^+(x_1^{-1}, \ldots, x_n^{-1}),$$

$$\delta^+(x) := \prod_{i=1}^n \frac{(x_i^2;q)_\infty}{(ax_i, bx_i, cx_i, dx_i;q)_\infty} \prod_{1\leq k<l\leq n} \frac{(x_kx_l^{-1}, x_kx_l;q)_\infty}{(tx_kx_l^{-1}, tx_kx_l;q)_\infty}.$$

Assume that $|a|, |b|, |c|, |d| \leq 1$ and that if a, b, c, and d are complex, then they appear in conjugate pairs. Assume furthermore that the pairwise products of a, b, c, and d are not equal to 1. Denote $du := du_1 \cdots du_n$ and $e^{iu} := (e^{iu_1}, \ldots, e^{iu_n})$. Suppose that $t \in (0,1)$; then

$$\langle f, g\rangle_{AW,t} := \int \cdot\cdot \int_{[-\pi,\pi]^n} f(e^{iu})\overline{g(e^{iu})}\delta(e^{iu};t)du, \quad f, g \in A^W,$$

is an Hermitian inner product on A^W. Define a partial order on P^+ in the following way: $\mu, \lambda \in P^+$. Then

$$\mu \leq \lambda \Leftrightarrow \sum_{j=1}^i \mu_j \leq \sum_{j=1}^i \lambda_j, \quad i = 1, \ldots, n. \tag{3.2}$$

Remark 3.1. For the root system $R = R^+ \cup (-R^+)$ of type BC_n, choose the positive roots R^+ by

$$R^+ = \{e_i\}_{i=1}^n \cup \{e_i \pm e_j\}_{1\leq i<j\leq n} \cup \{2e_i\}_{i=1}^n, \tag{3.3}$$

where $\{e_i\}_{i=1}^n$ is the standard orthonormal basis for $\mathbb{R}^n$, then P^+ coincides with the set of dominant weights, and $\lambda > \mu$ for $\lambda, \mu \in P^+$ iff $\lambda - \mu$ is a sum of positive roots (cf. [11]).

Definition 3.2. *Let $t \in (0,1)$. The Askey–Wilson polynomials*

$$\{Q_\lambda(x; a, b, c, d; q, t) \mid \lambda \in P^+\}$$

are defined by the following two conditions:

(1) $Q_\lambda(t) = \tilde{m}_\lambda + \sum_{\mu<\lambda;\mu\in P^+} c_{\lambda,\mu}(t)\tilde{m}_\mu$*, for certain* $c_{\lambda,\mu}(t) \in \mathbb{C}$*;*

(2) *if* $\mu < \lambda$ *and* $\mu \in P^+$*, then* $\langle Q_\lambda(t), \tilde{m}_\mu\rangle_{AW,t} = 0$.

Define a second-order q-difference operator $D_{AW,q,t}^{a,b,c,d}$ by

$$(D_{AW}f)(x) := \sum_{i=1}^n \big(\psi_i(x)(T_{q,i}f - f)(x) + \phi_i(x)\big(T_{q^{-1},i}f - f\big)(x)\big)$$

for $f \in A^W$, with

$$(T_{q,i}f)(x) := f(x_1, \ldots, x_{i-1}, qx_i, x_{i+1}, \ldots, x_n), \tag{3.4}$$

the q-shift in the ith component, and with $\psi_i(x; a, b, c, d; q, t)$ and $\phi_i(x; a, b, c, d; q, t)$ given by

$$\psi_i(x) := \frac{(1-ax_i)(1-bx_i)(1-cx_i)(1-dx_i)}{(1-x_i^2)(1-qx_i^2)} \prod_{l\neq i} \frac{(1-tx_ix_l)(1-tx_ix_l^{-1})}{(1-x_ix_l)(1-x_ix_l^{-1})}$$

$$\phi_i(x) := \psi_i(x_1^{-1}, \ldots, x_n^{-1}).$$

Koornwinder proved the following theorem in [11].

Theorem 3.3. *Let $t \in (0,1)$. Define $b_\lambda(a, b, c, d; q, t)$ for $\lambda \in P^+$ by*

$$b_\lambda := \sum_{j=1}^n \left(q^{-1}abcdt^{2n-j-1}(q^{\lambda_j} - 1) + t^{j-1}(q^{-\lambda_j} - 1)\right);$$

then $D_{AW,t}Q_\lambda(t) = b_\lambda(t)Q_\lambda(t)$ for all $\lambda \in P^+$ and $\langle Q_\lambda(t), Q_\mu(t)\rangle_{AW,t} = 0$ if $\lambda \neq \mu$.

For the one-variable case ($n = 1$), explicit expressions of the Askey–Wilson polynomials $\{Q_m(x; a, b, c, d; q) \mid m \in \mathbb{N}_0\}$ are given by

$$Q_m(x; a, b, c, d; q) = \frac{(ab, ac, ad; q)_m}{a^m\,(q^{m-1}abcd; q)_m}\,{}_4\phi_3\left[\begin{matrix} q^{-m}, q^{m-1}abcd, ax, ax^{-1} \\ ab, ac, ad \end{matrix}; q, q\right]$$

(cf. [4],[13]) and the following limit transitions hold:

$$\lim_{\epsilon\to 0}\left(\frac{\epsilon(cd)^{\frac12}}{q^{\frac12}}\right)^m Q_m\left(\frac{q^{\frac12}x}{\epsilon(cd)^{\frac12}}; \epsilon a(qd/c)^{\frac12}, \epsilon^{-1}(qc/d)^{\frac12}, -\epsilon^{-1}(qd/c)^{\frac12}, -\epsilon b(qc/d)^{\frac12}; q\right)$$
$$= P_m(x; a, b, c, d; q),$$

$$\lim_{\epsilon\to 0}\left(\frac{\epsilon}{q^{\frac12}}\right)^m Q_m\left(\frac{q^{\frac12}x}{\epsilon}; \epsilon q^{\frac12}b, \epsilon^{-1}q^{\frac12}, -q^{\frac12}, -q^{\frac12}a; q\right) = p_m(x; a, b; q).$$

(See [12, Propositions 6.1 and 6.2] and take into account that the Askey–Wilson polynomials used in those limit transitions are written as function of $(x+x^{-1})/2$ and that the polynomials used in [12] are not monic.)

The most obvious generalizations of these two limits to the n-variable case give two new second-order q-difference operators and a new set of eigenvalues.

Let $x=(x_1,\ldots,x_n)$ and denote $cx:=(cx_1,\ldots,cx_n)$ for $c\in\mathbb{C}$; then we have the following limits for the big q-Jacobi case:

$$\begin{aligned}\lim_{\epsilon\to 0}\psi_i&\left(\frac{q^{\frac12}x}{\epsilon(cd)^{\frac12}};\epsilon a(qd/c)^{\frac12},\epsilon^{-1}(qc/d)^{\frac12},-\epsilon^{-1}(qd/c)^{\frac12},-\epsilon b(qc/d)^{\frac12};q,t\right)\\&=h_i(x;a,b,c,d;q,t),\end{aligned}$$

with $h_i(x;a,b,c,d;q,t)$ given by

$$h_i(x;a,b,c,d;q,t):=qt^{n-1}\Big(a-\frac{c}{qx_i}\Big)\Big(b+\frac{d}{qx_i}\Big)\prod_{l\neq i}\frac{x_l-tx_i}{x_l-x_i},\tag{3.5}$$

$$\begin{aligned}\lim_{\epsilon\to 0}\phi_i&\left(\frac{q^{\frac12}x}{\epsilon(cd)^{\frac12}};\epsilon a(qd/c)^{\frac12},\epsilon^{-1}(qc/d)^{\frac12},-\epsilon^{-1}(qd/c)^{\frac12},-\epsilon b(qc/d)^{\frac12};q,t\right)\\&=g_i(x;c,d;q,t)\end{aligned}$$

with $g_i(x;c,d;q,t)$ given by

$$g_i(x;c,d;q,t):=\Big(1-\frac{c}{x_i}\Big)\Big(1+\frac{d}{x_i}\Big)\prod_{l\neq i}\frac{x_i-tx_l}{x_i-x_l},\tag{3.6}$$

and

$$\lim_{\epsilon\to 0}b_\lambda\left(\epsilon a(qd/c)^{\frac12},\epsilon^{-1}(qc/d)^{\frac12},-\epsilon^{-1}(qd/c)^{\frac12},-\epsilon b(qc/d)^{\frac12};q,t\right)=a_\lambda(a,b;q,t),$$

with

$$a_\lambda(a,b;q,t):=\sum_{j=1}^{n}\big(qabt^{2n-j-1}(q^{\lambda_j}-1)+t^{j-1}(q^{-\lambda_j}-1)\big).\tag{3.7}$$

For the little q-Jacobi case, we have the following limits:

$$\lim_{\epsilon\to 0}\psi_i\left(\frac{q^{\frac12}x}{\epsilon};\epsilon q^{\frac12}b,\epsilon^{-1}q^{\frac12},-q^{\frac12},-q^{\frac12}a;q,t\right)=h_i(x;b,a,1,0;q,t),$$

$$\lim_{\epsilon\to 0}\phi_i\left(\frac{q^{\frac12}x}{\epsilon};\epsilon q^{\frac12}b,\epsilon^{-1}q^{\frac12},-q^{\frac12},-q^{\frac12}a;q,t\right)=g_i(x;1,0;q,t),$$

and

$$\lim_{\epsilon\to 0}b_\lambda\left(\epsilon q^{\frac12}b,\epsilon^{-1}q^{\frac12},-q^{\frac12},-q^{\frac12}a;q,t\right)=a_\lambda(a,b;q,t).$$

Therefore, define the q-difference operator $D^{a,b,c,d}_{n,q,t}$ by

$$(D_n f)(x) := \sum_{j=1}^{n} \big(h_j(x)(T_{q,j}f - f)(x) + g_j(x)\big(T_{q^{-1},j}f - f\big)(x)\big); \tag{3.8}$$

then D_{AW} tends to $D^{a,b,c,d}_{n,q,t}$ (resp. to $D^{b,a,1,0}_{n,q,t}$) in the two limits we have just considered, and the eigenvalues $\{b_\lambda(a,b,c,d;q,t)\,|\,\lambda \in P^+\}$ tend to $\{a_\lambda(a,b;q,t)\,|\,\lambda \in P^+\}$. For $n=1$, $D^{a,b,c,d}_{1,q,t}$ and $D^{b,a,1,0}_{1,q,t}$ correspond with the second-order q-difference operators for which the one-variable big (resp. little) q-Jacobi polynomials are joint eigenfunctions ((2.4) and (2.5) (resp. (2.9))), and $\{a_m(a,b;q,t)\,|\,m \in \mathbb{N}_0\}$ is exactly the corresponding set of eigenvalues (formula (2.6)). Therefore, we denote

$$D^{a,b,c,d}_{B,q,t} := D^{a,b,c,d}_{n,q,t} \tag{3.9}$$

and

$$D^{a,b}_{L,q,t} := D^{b,a,1,0}_{n,q,t}. \tag{3.10}$$

In section 5, we will see that the multivariable big (resp. little) q-Jacobi polynomials are joint eigenfunctions of D_B (resp. D_L) with eigenvalues $\{a_\lambda\,|\,\lambda \in P^+\}$. In [18], it is shown that the formal limit transitions of the second-order q-difference operator D_{AW} that we discussed in this section can be used to prove limit transitions from multivariable Askey–Wilson polynomials to multivariable big and little q-Jacobi polynomials.

Remark 3.4. Van Diejen mentioned similar limit transitions in [5] but did not look for eigenfunctions of the newly obtained q-difference operators. In his terminology, the limit transitions correspond to sending the center of mass in an n-particle-difference Calogero–Moser system with trigonometric potentials (Hamiltonian given by D_{AW}) to infinity.

4. Triangularity of the second-order q-difference operator $D^{a,b,c,d}_{n,q,t}$.

Note that

$$\Delta(x)^{-1}(T_{t,i}\Delta)(x) = \prod_{l\neq i} \frac{x_l - tx_i}{x_l - x_i}, \tag{4.1}$$

where $\Delta(x)$ is the Vandermonde determinant $\Delta(x) := \prod_{1\le i<j\le n}(x_i - x_j)$. Therefore, we can rewrite the second-order q-difference operator D_n (given by (3.8)) in the following form:

$$(D_n f)(x) = \Delta(x)^{-1}(\tilde{D}_n f)(x), \tag{4.2}$$

with

$$(\tilde{D}_n f)(x) := \sum_{i=1}^{n} \left(\tilde{h}_i(x)(T_{q,i}f - f)(x) + \tilde{g}_i(x)\big(T_{q^{-1},i}f - f\big)(x)\right), \tag{4.3}$$

$$\tilde{h}_i(x) = q\left(a - \frac{c}{qx_i}\right)\left(b + \frac{d}{qx_i}\right)t^{n-1}(T_{t,i}\Delta)(x), \tag{4.4}$$

$$\tilde{g}_i(x) = \left(1 - \frac{c}{x_i}\right)\left(1 + \frac{d}{x_i}\right)t^{n-1}\big(T_{t^{-1},i}\Delta\big)(x). \tag{4.5}$$

Denote $\mathbb{C}[x_1,\ldots,x_n]$ for the $\mathbb{C}$-algebra of polynomials in the variables $x_1,\ldots,x_n$. We have the following result.

Lemma 4.1.

$$\tilde{D}_n\left(\mathbb{C}[x_1,\ldots,x_n]\right)\subseteq\mathbb{C}[x_1,\ldots,x_n].$$

Proof. For $f\in\mathbb{C}[x_1,\ldots,x_n]$, define the backward partial q-derivative in the ith coordinate by

$$\left(D_q^{i,-}f\right)(x):=\frac{(f-T_{q,i}f)(x)}{(1-q)x_i}.$$

Note that $D_q^{i,-}$ maps $\mathbb{C}[x_1,\ldots,x_n]$ into itself. Now it can easily be checked that

$$(\tilde{D}_nf)(x)=\sum_{j=1}^{n}\left(A_j(x)\left(T_{q^{-1},j}\left(\left(D_q^{j,-}\right)^2f\right)\right)(x)+B_j(x)\left(T_{q^{-1},j}\left(D_q^{j,-}f\right)\right)(x)\right),$$

with

$$A_j(x)=(1-q)^2q^{-2}(qax_j-c)(qbx_j+d)t^{n-1}(T_{t,j}\Delta)(x),$$

$$\begin{aligned}B_j(x)=\frac{(1-q)}{q}t^{n-1}\Bigg(&\left(x_j+(d-c)-\frac{cd}{x_j}\right)(T_{t^{-1},j}\Delta)(x)\\&-\left(q^2abx_j+(qad-qbc)-\frac{cd}{x_j}\right)(T_{t,j}\Delta)(x)\Bigg).\end{aligned}$$

The lemma follows because $\left((T_{t,j}-T_{t^{-1},j})\Delta\right)(x)\in\mathbb{C}[x_1,\ldots,x_n]$ is divisible by x_j in $\mathbb{C}[x_1,\ldots,x_n]$. □

Let S_n be the permutation group of $\{1,\ldots,n\}$. S_n acts on $\mathbb{C}[x_1,\ldots,x_n]$ by permutation of the variables $x_1,\ldots,x_n$. Denote $\mathbb{C}[x_1,\ldots,x_n]^{S_n}$ for the subalgebra (over $\mathbb{C}$) of symmetric polynomials.

For $\lambda\in P^+$, define the symmetric monomial function m_λ by $m_\lambda(x):=\sum_{\mu\in S_n\lambda}x^\mu$, with $x^\mu:=x_1^{\mu_1}\ldots x_n^{\mu_n}$ and $w\lambda:=(\lambda_{w^{-1}(1)},\ldots,\lambda_{w^{-1}(n)})$. Then $\{m_\lambda\,|\,\lambda\in P^+\}$ is a $\mathbb{C}$-basis for $\mathbb{C}[x_1,\ldots,x_n]^{S_n}$.

A second basis is given by the Schur functions $\{s_\lambda\,|\,\lambda\in P^+\}$, where

$$s_\lambda(x):=\Delta(x)^{-1}\sum_{w\in S_n}\det(w)x^{w(\lambda+\delta)},$$

where $\det(w)$ is the determinant of the linear map $w:\mathbb{R}^n\to\mathbb{R}^n$ given by $w(e_i):=e_{w(i)}$ $(i=1,\ldots,n)$ for an arbitrary basis $\{e_1,\ldots,e_n\}$ of $\mathbb{R}^n$, and where

$$\delta:=(n-1,n-2,\ldots,1,0)\in P^+. \tag{4.6}$$

Let $\lambda\in P^+$; then

$$s_\lambda=m_\lambda+\sum_{\mu<\lambda;\mu\in P^+}c_{\lambda,\mu}m_\mu \tag{4.7}$$

for certain $c_{\lambda,\mu}\in\mathbb{R}$. See [17] for more details about Schur functions.

PROPOSITION 4.2. *Let $a, b, c, d, t \in \mathbb{C}$ and $\lambda \in P^+$. Then*

$$D_n m_\lambda = a_\lambda m_\lambda + \sum_{\mu<\lambda;\mu\in P^+} d_{\lambda,\mu} m_\mu$$

for certain $d_{\lambda,\mu} \in \mathbb{C}$, with $a_\lambda = a_\lambda(a, b; q, t)$ given by (3.7). *$d_{\lambda,\mu}$ and a_λ depend polynomially on $a, b, c, d,$ and t.*

Proof. Let $e_1, \ldots, e_n$ be the standard basis of $\mathbb{R}^n$ and $\langle\, .\,,\, .\,\rangle$ be the standard inner product on $\mathbb{R}^n$. Let S_n act on $\mathbb{R}^n$ by permutation of the basis $\{e_1, \cdots, e_n\}$. Define

$$\tilde{P} := \{\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{Z}^n \,|\, \lambda_1 \geq \cdots \geq \lambda_n\},$$

and give $\tilde{P}$ the same partial order as P^+ (see (3.2)). For $\mu \in \mathbb{Z}^n$, define

$$J_\mu := \sum_{w\in S_n} \det(w) x^{w\mu}.$$

Then $J_\mu = 0$ unless $\mu = w(\nu + \delta)$ for certain $w \in S_n$ and $\nu \in \tilde{P}$, and in that case, we have $J_\mu = \det(w) J_{\nu+\delta}$. Write $\tilde{D}_n = \psi_+ + \psi_-$ with

$$(\psi_+ f)(x) := \sum_{j=1}^{n} \tilde{h}_j(x)(T_{q,j} f - f)(x), \qquad (\psi_- f)(x) := \sum_{j=1}^{n} \tilde{g}_j(x)(T_{q^{-1},j} f - f)(x).$$

Let $S_n^\lambda \subseteq S_n$ be the stabilizer of $\lambda \in P^+$, so $S_n^\lambda := \{w \in S_n \,|\, w\lambda = \lambda\}$. Denote $T_{q,e_i} := T_{q,i}$ and $\sigma := e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^n$. Using the fact that

$$\Delta(x) = \sum_{w\in S_n} \det(w) x^{w\delta}$$

and

$$m_\lambda(x) = |S_n^\lambda|^{-1} \sum_{v\in S_n} x^{v\lambda} \quad (\lambda \in P^+),$$

we obtain, for $\lambda \in P^+$,

$$\begin{aligned}
(\psi_+ m_\lambda)(x) &= \frac{1}{(n-1)!} \sum_{u\in S_n} q t^{n-1} \left(a - \frac{c}{q} x^{-u\sigma}\right) \left(b + \frac{d}{q} x^{-u\sigma}\right) (T_{t,u\sigma}\Delta)(x) \\
&\quad \times (T_{q,u\sigma} m_\lambda - m_\lambda)(x) \\
&= \frac{1}{(n-1)!|S_n^\lambda|} \sum_{u,v,w\in S_n} q t^{n-1} \left(a - \frac{c}{q} x^{-u\sigma}\right) \left(b + \frac{d}{q} x^{-u\sigma}\right) \det(w) \\
&\quad \times t^{\langle u\sigma, w\delta\rangle} (q^{\langle u\sigma, v\lambda\rangle} - 1) x^{w\delta + v\lambda} \\
&= \frac{1}{(n-1)!|S_n^\lambda|} \sum_{u',v'\in S_n} t^{\langle \sigma, v'\delta+\delta\rangle} (q^{\langle \sigma, v'u'\lambda\rangle} - 1) \left(qab \sum_{w\in S_n} \det(w) x^{w(\delta+u'\lambda)} \right. \\
&\left. + (ad - bc) \sum_{w\in S_n} \det(w) x^{w(\delta+u'\lambda-v'^{-1}\sigma)} - \frac{cd}{q} \sum_{w\in S_n} \det(w) x^{w(\delta+u'\lambda-2v'^{-1}\sigma)} \right).
\end{aligned}$$

The third equality is obtained by the substitution of $u' = w^{-1}v$ and $v' = u^{-1}w$. Similarly, we have

$$(\psi_- m_\lambda)(x) = \frac{1}{(n-1)!|S_n^\lambda|}\sum_{u',v'\in S_n} t^{\langle\sigma,\delta-v'\delta\rangle}(q^{-\langle\sigma,v'u'\lambda\rangle}-1)\Bigg(\sum_{w\in S_n}\det(w)x^{w(\delta+u'\lambda)}$$
$$+(d-c)\sum_{w\in S_n}\det(w)x^{w(\delta+u'\lambda-v'^{-1}\sigma)} - cd\sum_{w\in S_n}\det(w)x^{w(\delta+u'\lambda-2v'^{-1}\sigma)}\Bigg).$$

Let $w,u',v'\in S_n$ and $\lambda\in P^+$; then $w(\delta+u'\lambda)\le\delta+wu'\lambda\le\delta+\lambda$, and

$$w(\delta+u'\lambda)=\delta+\lambda \Leftrightarrow w=(1) \quad\text{and}\quad u'\in S_n^\lambda.$$

Furthermore, we have $w(\delta+u'\lambda-v'^{-1}\sigma)\le\delta+wu'\lambda-wv'^{-1}\sigma<\delta+wu'\lambda\le\delta+\lambda$ and $w(\delta+u'\lambda-2v'^{-1}\sigma)<\delta+\lambda$. Thus

$$\psi_\epsilon m_\lambda = \alpha_\lambda^\epsilon J_{\lambda+\delta} + \sum_{\mu<\lambda;\mu\in\tilde{P}}\beta_{\lambda,\mu}^\epsilon J_{\mu+\delta}, \quad \epsilon=\pm,$$

with $\beta_{\mu,\lambda}^\epsilon\in\mathbb{C}$, $\alpha_\lambda^+=\sum_{i=1}^n qabt^{2n-i-1}(q^{\lambda_i}-1)$, and $\alpha_\lambda^-=\sum_{i=1}^n t^{i-1}(q^{-\lambda_i}-1)$. Lemma 4.1 implies

$$\tilde{D}_n m_\lambda = \psi_+ m_\lambda + \psi_- m_\lambda = a_\lambda J_{\lambda+\delta} + \sum_{\mu<\lambda;\mu\in P^+} c_{\lambda,\mu}J_{\mu+\delta} \quad (\lambda\in P^+)$$

for certain $c_{\lambda,\mu}\in\mathbb{C}$. Formula (4.7) gives now the triangularity property.

Finally, note that the coefficients of J_ν ($\nu\in P^+$) in the expressions for $\psi_+ m_\lambda$ and $\psi_- m_\lambda$ depend polynomially on a,b,c,d, and t. Therefore, the coefficients of $D_n m_\lambda$ with respect to the basis of monomial symmetric functions depend polynomially on a,b,c,d, and t. □

5. Multivariable big and little *q*-Jacobi polynomials. The $\mathbb{R}$-algebra of symmetric polynomials in $x_1,\dots,x_n$ will be denoted by $\mathcal{A}^{\mathcal{S}}$, so $\mathcal{A}^{\mathcal{S}}:=\mathbb{R}[x_1,\dots,x_n]^{S_n}$. We first define inner products $\langle\,.\,,\,.\,\rangle_{B,n,q,t}^{a,b,c,d}$ and $\langle\,.\,,\,.\,\rangle_{L,n,q,t}^{a,b}$ on $\mathcal{A}^{\mathcal{S}}$, which generalize the inner products $\langle\,.\,,\,.\,\rangle_{B,1}$ and $\langle\,.\,,\,.\,\rangle_{L,1}$ ((2.3) and (2.8)) to the multivariable case.

For the big q-Jacobi case, we fix some $(a,b,c,d)\in V_B^q$ unless otherwise stated (V_B^q is defined in section 2). Define a symmetric bilinear form $\langle\,.\,,\,.\,\rangle_{B,n,q,t}^{a,b,c,d}$ for $t\in(0,1)$ on $\mathcal{A}^{\mathcal{S}}$ by

$$\langle f,g\rangle_{B,t} := \sum_{j=0}^n \langle f,g\rangle_{j,B,t}, \quad f,g\in\mathcal{A}^{\mathcal{S}}, \tag{5.1}$$

with $\langle f,g\rangle_{j,B,t}$ given by the following multidimensional Jackson integral:

$$\int_{x_1=0}^{c}\int_{x_2=0}^{tx_1}\cdots\int_{x_j=0}^{tx_{j-1}}\int_{x_{j+1}=-dt^{n-j-1}}^{0}\int_{x_{j+2}=-dt^{n-j-2}}^{qt^{-1}x_{j+1}} \tag{5.2}$$
$$\cdots\int_{x_n=-d}^{qt^{-1}x_{n-1}} f(x)g(x)w_j(x;t)d_qx,$$

with $d_qx := d_qx_n \cdots d_qx_1$ and the weight function $w_j(x;a,b,c,d;q,t)$ given by

$$w_j(x;t) := d_j^\tau \left(\prod_{i=1}^{n} \frac{(qx_i/c, -qx_i/d;q)_\infty}{(qax_i/c, -qbx_i/d;q)_\infty} \right) \Delta_\tau^j(x), \tag{5.3}$$

with $t = q^\tau$ and

$$\begin{aligned} \Delta_\tau^j(x) := \Delta(x) &\left(\prod_{\substack{1 \le k < m \le n \\ k \le j}} |x_k|^{2\tau-1} \left(q^{1-\tau} \frac{x_m}{x_k}; q \right)_{2\tau-1} \right) \\ &\times \prod_{j<k<m\le n} |x_m|^{2\tau-1} \left(q^{1-\tau} \frac{x_k}{x_m}; q \right)_{2\tau-1} \end{aligned} \tag{5.4}$$

and with $d_j^\tau = d_j^\tau(c,d)$ a positive constant given by

$$d_j^\tau := \prod_{\substack{1 \le k < m \le n \\ k \le j}} |y_{mk}|^{2\tau-1} \frac{(q^{1-\tau} y_{mk}^{-1};q)_{2\tau-1}}{(q^{1-\tau} y_{mk};q)_{2\tau-1}}, \quad y_{mk} := \frac{-d}{c} q^{(n-m-k+1)\tau}. \tag{5.5}$$

In view of (2.1), we have that the measure associated with $\langle \, . \, , \, . \, \rangle_{j,B}$ has infinitely many discrete mass points given by the set

$$\begin{aligned} W_B^j := \{(x_1, \ldots, x_n) \,|\, & x_i = cq^{(i-1)\tau + k_i} \text{ if } i \le j \text{ and } 0 \le k_1 \le \cdots \le k_j, \\ & x_i = -dq^{(n-i)\tau + k_i} \text{ if } i > j \text{ and } 0 \le k_n \le \cdots \le k_{j+1}\}. \end{aligned} \tag{5.6}$$

We have that $w_j(x;t) > 0$ for all $x \in W_B^j(t)$ and all $t \in (0,1)$. Indeed, we only need to check that Δ_τ^j is positive because $(a,b,c,d) \in V_B^q$. For Δ_τ^j, it is easily checked that the terms of the form

$$\left(q^{1-\tau} x_p/x_r;q\right)_{2\tau-1} = \frac{\left(q^{1-\tau} x_p/x_r;q\right)_\infty}{\left(q^{\tau} x_p/x_r;q\right)_\infty}$$

are positive on the mass points since both the numerator and the denominator are positive on the mass points. Furthermore, note that for $x \in W_B^j$ we have the inequalities $x_1 > \cdots > x_n$, so the Vandermonde determinant $\Delta(x)$ is positive for $x \in W_B^j$. Finally, it can be shown that w_j is bounded on W_B^j (in fact, we will see in the proof of Proposition 5.5 (section 6) that $w_j(x;q^\tau)$ is uniformly bounded on the set $\{(x,\tau) \,|\, \tau \in K, x \in W_B^j(q^\tau)\}$, with K an arbitrary compact subset of $(0,\infty)$), so $\langle \, . \, , \, . \, \rangle_{B,t}$ is a well-defined positive definite inner product for all $t \in (0,1)$.

For the little q-Jacobi case, we fix $(a,b) \in V_L^q$ unless otherwise stated (V_L^q is defined in section 2). Define a symmetric bilinear form $\langle \, . \, , \, . \, \rangle_{L,n,q,t}^{a,b}$ on $\mathcal{A}^\mathcal{S}$ by

$$\langle f,g \rangle_{L,t} := \int_{x_1=0}^{1} \int_{x_2=0}^{tx_1} \cdots \int_{x_n=0}^{tx_{n-1}} f(x)g(x)v(x;t) d_qx, \quad f,g \in \mathcal{A}^\mathcal{S}, \tag{5.7}$$

with the weight function $v(x;a,b;q,t)$ given by

$$v(x;t) := \left(\prod_{i=1}^{n} \frac{(qx_i;q)_\infty}{(qbx_i;q)_\infty} x_i^\alpha \right) \Delta_\tau(x) \quad (a = q^\alpha, t = q^\tau), \tag{5.8}$$

$$\Delta_\tau(x) := \Delta(x) \prod_{1 \le i < j \le n} |x_i|^{2\tau-1} \left(q^{1-\tau} \frac{x_j}{x_i}; q \right)_{2\tau-1}. \tag{5.9}$$

The measure associated with the inner product $\langle\, .\, ,\, .\,\rangle_L$ has infinitely many discrete mass points given by the set

$$W_L := \{(x_1, \ldots, x_n) \,|\, x_i = q^{(i-1)\tau + k_i} \text{ with } 0 \leq k_1 \leq \cdots \leq k_n\}. \tag{5.10}$$

By a similar argument as in the big q-Jacobi case, we have that $v(x;t) > 0$ for all $x \in W_L(t)$ and all $t \in (0,1)$ because $(a,b) \in V_L^q$. Furthermore, $v(x)(\prod_{i=1}^n x_i^{-\alpha})$ is bounded on W_L (we will see in the proof of Proposition 5.5 (section 6) that $v(x;q^\tau)(\prod_{i=1}^n x_i^{-\alpha})$ is uniformly bounded on $\{(x,\tau) \,|\, \tau \in K, x \in W_L(q^\tau)\}$, with K an arbitrary compact subset of $(0,\infty)$), so $\langle\, .\, ,\, .\,\rangle_{L,t}$ is well defined because $\alpha > -1$ and positive definite for all $t \in (0,1)$.

Definition 5.1. *Let $t \in (0,1)$. The big q-Jacobi polynomials*

$$\{P_\lambda^B(\,.\,; a, b, c, d; q, t) \,|\, \lambda \in P^+\}$$

are defined by the following two conditions: let $\lambda \in P^+$; then

(1) $P_\lambda^B(t) = m_\lambda + \sum_{\mu < \lambda; \mu \in P^+} c_{\lambda,\mu}(t) m_\mu$ *for some* $c_{\lambda,\mu}(t) \in \mathbb{R}$*;*

(2) $\langle P_\lambda^B(t)\,,\, m_\mu\rangle_{B,t} = 0$ *if* $\mu < \lambda$, $\mu \in P^+$.

Definition 5.2. *Let $t \in (0,1)$. The little q-Jacobi polynomials*

$$\{P_\lambda^L(\,.\,; a, b; q, t) \,|\, \lambda \in P^+\}$$

are defined by the following two conditions: let $\lambda \in P^+$; then

(1) $P_\lambda^L(t) = m_\lambda + \sum_{\mu < \lambda; \mu \in P^+} d_{\lambda,\mu}(t) m_\mu$ *for some* $d_{\lambda,\mu}(t) \in \mathbb{R}$*;*

(2) $\langle\, P_\lambda^L(t), m_\mu\,\rangle_{L,t} = 0$ *if* $\mu < \lambda$, $\mu \in P^+$.

For $n = 1$, the inner products $\langle\, .\, ,\, .\,\rangle_B$ and $\langle\, .\, ,\, .\,\rangle_L$ are the same as the inner products given by (2.3) and (2.8), respectively. Thus for $n = 1$, the big (resp. little) q-Jacobi polynomials given by Definition 5.1 (resp. Definition 5.2) are exactly the one-variable big (resp. little) q-Jacobi polynomials as defined in section 2 (Definition 2.1 (resp. Definition 2.2)).

Observe that the multivariable big q-Jacobi polynomials depend (apart from q) only on a, b, t, and the ratio c/d. Indeed, let $f \in \mathcal{A}^{\mathcal{S}}$ and define $f_d \in \mathcal{A}^{\mathcal{S}}$ by $f_d(x) := f(dx)$; then

$$\langle f, g\rangle_{B,n,q,t}^{a,b,c,d} = d^{2\tau\binom{n}{2}+n} \langle f_d, g_d\rangle_{B,n,q,t}^{a,b,\frac{c}{d},1}, \quad f, g \in \mathcal{A}^{\mathcal{S}},$$

because

$$\int_0^\alpha h(u) d_q u = \alpha \int_0^1 h(\alpha u) d_q u \quad (\alpha \neq 0), \tag{5.11}$$

and $d_j^\tau(c,d) = d_j^\tau(c/d, 1)$, so $w_j(dx; a, b, c, d; q, t) = d^{2\tau\binom{n}{2}} w_j(x; a, b, c/d, 1; q, t)$. Therefore, we have that

$$d^{-|\lambda|} P_\lambda^B(dx; a, b, c, d; q, t) = P_\lambda^B\left(x; a, b, \frac{c}{d}, 1; q, t\right), \quad \lambda \in P^+,$$

where $|\lambda| := \sum_{i=1}^n \lambda_i$.

Remark 5.3. If we compare the weight functions w_j with the function w given by

$$w(x) := \left(\prod_{i=1}^n \frac{(qx_i/c, -qx_i/d; q)_\infty}{(qax_i/c, -qbx_i/d; q)_\infty}\right) \tilde{\Delta}_\tau(x), \tag{5.12}$$

with

$$\tilde{\Delta}_\tau(x) := \Delta(x) \prod_{1\le i<j\le n} \operatorname{sgn}(x_i)|x_i|^{2\tau-1}\left(q^{1-\tau}\frac{x_j}{x_i};q\right)_{2\tau-1}, \tag{5.13}$$

and $\operatorname{sgn}(x_i)=1$ if $x_i\ge 0$ and $=-1$ if $x_i<0$, then we have that

$$\Delta_\tau^j(x) = \left(\prod_{k=1}^{j}\operatorname{sgn}(x_k)^{n-k}\right)\left(\prod_{j<k<m\le n}\operatorname{sgn}(x_k)\psi_\tau\left(\frac{x_m}{x_k}\right)\right)\tilde{\Delta}_\tau(x) \tag{5.14}$$

with the function ψ_τ given by

$$\psi_\tau(z) := |z|^{2\tau-1}\frac{\left(q^{1-\tau}z^{-1};q\right)_{2\tau-1}}{\left(q^{1-\tau}z;q\right)_{2\tau-1}}. \tag{5.15}$$

ψ_τ is a quasi-constant function, i.e., $\psi_\tau(qz)=\psi_\tau(z)$, so $w_j(x)=\phi_j(x)w(x)$ for some quasi-constant function ϕ_j ($T_{q,i}\phi_j=\phi_j$ for all i). The essential difference between $w(x)$ and $w_j(x)$ on W_B^j is that $w(x)$ can have poles on W_B^j, while $w_j(x)$ has no poles on W_B^j. Therefore, one can think of $\langle\,.\,,\,.\,\rangle_B$ as q-integration over the set of mass points $\cup_{j=0}^n W_B^j$ with respect to the weight function w, whereby one should resolve the poles of w on W_B^j when q-integrating over W_B^j by slightly modifying the weight function w (i.e., multiplying w with the quasi-constant function $d_j\phi_j$). The constants d_j in the definition of $w_j(x)$ will turn out to be crucial for the self-adjointness of D_B with respect to $\langle\,.\,,\,.\,\rangle_B$. Note that

$$d_j^\tau = \prod_{\substack{1\le k<m\le n\\ k\le j}}\psi_\tau(y_{mk}),\quad \text{with } y_{mk} := \frac{-d}{c}q^{(n-m-k+1)\tau}, \tag{5.16}$$

so d_j^τ can also be expressed in terms of the quasi-constant function ψ_τ.

The inner products simplify when $\tau=k\in\mathbb{N}$. In that case, we have that $w_j(x)=w(x)$ on W_B^j for $j=0,\ldots,n$ (with $w(x)$ given by (5.12)). This follows from (5.14) and (5.16) since $\psi_k(z)=-\operatorname{sgn}(z)$. Furthermore, it holds that

$$\tilde{\Delta}_k(x) = (-1)^{k\binom{n}{2}}q^{-\binom{k}{2}\binom{n}{2}}\prod_{l=0}^{k-1}\prod_{i\ne j}(x_i-q^lx_j), \tag{5.17}$$

so $\tilde{\Delta}_k(x)$ is symmetric, and $\tilde{\Delta}_k(x)=0$ if $x_i=q^lx_j$ for certain $i\ne j$ and certain $l\in\{0,\ldots,k-1\}$. Thus when $\tau=k\in\mathbb{N}$, we have, for $f,g\in\mathcal{A}^{\mathcal{S}}$,

$$\begin{aligned}\langle f,g\rangle_{B,q^k} &= \int_{x_1=-d}^{c}\int_{x_2=-d}^{x_1}\cdots\int_{x_n=-d}^{x_{n-1}} f(x)g(x)w(x;q^k)d_qx,\\ &= \frac{1}{n!}\int_{x_1=-d}^{c}\cdots\int_{x_n=-d}^{c} f(x)g(x)w(x;q^k)d_qx,\end{aligned} \tag{5.18}$$

$$\begin{aligned}\langle f,g\rangle_{L,q^k} &= \int_{x_1=0}^{1}\int_{x_2=0}^{x_1}\cdots\int_{x_n=0}^{x_{n-1}} f(x)g(x)v(x;q^k)d_qx\\ &= \frac{1}{n!}\int_{x_1=0}^{1}\cdots\int_{x_n=0}^{1} f(x)g(x)v(x;q^k)d_qx.\end{aligned} \tag{5.19}$$

Here we have used that the weight functions $w(x;q^k)$ and $v(x;q^k)$ are zero for $x=(x_1,\dots,x_n)$ with $x_i=x_j$ for some $1\leq i\neq j\leq n$, so if $x\in\mathbb{R}^n$ contributes to the support of the orthogonality measure, then the S_n orbit of $x\in\mathbb{R}^n$ has cardinality $n!$.

Finally, we may replace $w(x;a,b,c,d;q,q^k)$ in (5.18) by $\tilde{w}(x;a,b,c,d;q,q^k)$,

$$\tilde{w}(x)=\frac{n!}{\Gamma_{q^k}(n+1)}\left(\prod_{i=1}^{n}w_B(x_i)\right)\prod_{1\leq i<j\leq n}x_i^{2k}\left(q^{1-k}\frac{x_j}{x_i};q\right)_{2k},$$

and $v(x;a,b;q,q^k)$ in (5.19) by $\tilde{v}(x;a,b;q,q^k)$,

$$\tilde{v}(x)=\frac{n!}{\Gamma_{q^k}(n+1)}\left(\prod_{i=1}^{n}v_L(x_i)\right)\prod_{1\leq i<j\leq n}x_i^{2k}\left(q^{1-k}\frac{x_j}{x_i};q\right)_{2k},$$

with the q-gamma function $\Gamma_q(a)$ $(a\notin-\mathbb{N}_0)$ defined by

$$\Gamma_q(a):=\frac{(q;q)_{a-1}}{(1-q)^{a-1}},$$

because w and v are symmetric functions such that

$$\tilde{w}(x)=\frac{n!}{\Gamma_{q^k}(n+1)}w(x)\prod_{i<j}\frac{x_i-q^kx_j}{x_i-x_j},\qquad \tilde{v}(x)=\frac{n!}{\Gamma_{q^k}(n+1)}v(x)\prod_{i<j}\frac{x_i-q^kx_j}{x_i-x_j},$$

and

$$\sum_{w\in S_n}\prod_{i<j}\frac{x_{w(i)}-q^kx_{w(j)}}{x_{w(i)}-x_{w(j)}}=\Gamma_{q^k}(n+1)$$

(cf. [8, p. 1479]).

Remark 5.4. For $t=q^k$, $k\in\mathbb{N}$, $\langle 1,1\rangle_L$ and $\langle 1,1\rangle_B$ are (up to the constant $1/\Gamma_{q^k}(n{+}1)$) the q-extensions of Selberg's multidimensional beta-integrals, introduced by Askey [3]. Askey's conjectured evaluations of these multidimensional q-integrals have recently been proved.

Let $t=q^k$, $k\in\mathbb{N}$, $a=q^\alpha$, and $b=q^\beta$; then Habsieger [8] and Kadell [10] have independently proved that

$$\langle 1,1\rangle_L=q^{k(\alpha+1)\binom{n}{2}+2k^2\binom{n}{3}}\prod_{j=1}^{n}\frac{\Gamma_q(\alpha+1+(j-1)k)\Gamma_q(\beta+1+(j-1)k)\Gamma_q(jk)}{\Gamma_q(\alpha+\beta+2+(n+j-2)k)\Gamma_q(k)},$$

and Evans [6] has proved that

$$\begin{aligned}\langle 1,1\rangle_B=&q^{k^2\binom{n}{3}-\binom{k}{2}\binom{n}{2}}\prod_{j=1}^{n}\Big(\frac{\Gamma_q(\alpha+1+(j-1)k)\Gamma_q(\beta+1+(j-1)k)\Gamma_q(jk)}{\Gamma_q(\alpha+\beta+2+(n+j-2)k)\Gamma_q(k)}\\&\times\frac{(-d/c;q)_\infty(-c/d;q)_\infty(cd)^{1+(j-1)k}}{\big((-d/c)q^{\alpha+1+(j-1)k};q\big)_\infty\big((-c/d)q^{\beta+1+(j-1)k};q\big)_\infty(c+d)}\Big).\end{aligned}$$

The inner products $\langle\,.\,,\,.\,\rangle_{B,t}$ and $\langle\,.\,,\,.\,\rangle_{L,t}$ depend continuously on $t\in(0,1)$, in the following sense.

Proposition 5.5. *Let $f, g \in \mathcal{A}^{\mathcal{S}}$.*

(1) *$\langle f, g\rangle_{B,t}$ is continuous in t for $t \in (0,1)$.*

(2) *$\langle f, g\rangle_{L,t}$ is continuous in t for $t \in (0,1)$.*

We will omit the proof of the proposition in this section because it is rather long and technical. The proof will be given in section 6 (Proposition 6.1).

Let $\lambda \in P^+$. It is clear from the proof of Proposition 4.2, that the coefficients in the expansion of the symmetric polynomial $D^{a,b,c,d}_{n,q,t} m_\lambda$ with respect to the basis of monomials $\{m_\mu \,|\, \mu \in P^+\}$ are real for $(a,b,c,d) \in V^q_B$. Therefore, $D_B m_\lambda \in \mathcal{A}^{\mathcal{S}}$. Similarly, we have that $D_L m_\lambda \in \mathcal{A}^{\mathcal{S}}$.

Theorem 5.6. *Let $t \in (0,1)$.*

(1) *$D_{B,t}$ is self-adjoint with respect to $\langle\,.\,,\,.\,\rangle_{B,t}$.*

(2) *$D_{L,t}$ is self-adjoint with respect to $\langle\,.\,,\,.\,\rangle_{L,t}$.*

The proof will be given in section 6 (Theorem 6.5). The two essential ingredients for the proof are a special version of the q-partial integration rule (Lemma 6.4) and certain functional relations for the weight functions (Proposition 6.3). Self-adjointness is then a consequence of the fact that stock terms (which come from the q-partial integration rule) are zero or cancel. In the big q-Jacobi case, the specific positive constants d_j in the weight functions w_j are crucial for the cancellation of certain stock terms.

With the aid of Propositions 4.2 and 5.5 and Theorem 5.6, it is now straightforward to proof the main theorem. The proof is similar to proofs given by Macdonald in [15] and [16] (see also the second edition of [17]), and Koornwinder in [11].

Theorem 5.7. *Let $t \in (0,1)$.*

(1) *Let $\lambda \in P^+$; then*

$$D_{B,t} P^B_\lambda(t) = a_\lambda(t) P^B_\lambda(t),$$

with a_λ given by (3.7). *For $\lambda, \mu \in P^+$, we have*

$$\langle P^B_\lambda(t), P^B_\mu(t)\rangle_{B,t} = 0 \qquad \textit{if } \lambda \neq \mu.$$

(2) *Let $\lambda \in P^+$; then*

$$D_{L,t} P^L_\lambda(t) = a_\lambda(t) P^L_\lambda(t).$$

For $\lambda, \mu \in P^+$, we have

$$\langle P^L_\lambda(t), P^L_\mu(t)\rangle_{L,t} = 0 \qquad \text{if } \lambda \neq \mu.$$

Proof. (1) Proposition 4.2 and Theorem 5.6 imply that $P^B_\lambda(t)$ is an eigenfunction of $D_{B,t}$ with eigenvalue $a_\lambda(t)$. Fix $(a,b,c,d) \in V^q_B$, fix $\mu, \lambda \in P^+$, $\mu \neq \lambda$, and fix $t \in (0,1)$ such that $a_\lambda(a,b;q,t) \neq a_\mu(a,b;q,t)$. The self-adjointness of $D_{B,t}$ then gives that $\langle P^B_\lambda(t), P^B_\mu(t)\rangle_{B,t} = 0$. Note that $a_\lambda(a,b;q,t) \in \mathbb{R}[t]$ and $a_\lambda(a,b;q,t) = a_\mu(a,b;q,t)$ as polynomials in t iff $\lambda = \mu$ because $ab \notin \{q^{-2}, q^{-3}, \ldots\}$. Therefore, (1) will be proved if we prove that $\langle P^B_\lambda(t), P^B_\mu(t)\rangle_{B,t}$ is continuous in t for $t \in (0,1)$. This follows from Proposition 5.5. The proof of (2) is similar. □

For $t = 1$, define

$$\langle f, g\rangle_{B,n,1} := \lim_{t\uparrow 1} \langle f, g\rangle_{B,n,t}, \tag{5.20}$$

$$\langle f, g\rangle_{L,n,1} := \lim_{t\uparrow 1} \langle f, g\rangle_{L,n,t} \tag{5.21}$$

for $f, g \in \mathcal{A}^{\mathcal{S}}$, provided that the limits exist. We then have the following proposition.

Proposition 5.8. *$\langle\,.\,,\,.\,\rangle_{B,n,1}$ and $\langle\,.\,,\,.\,\rangle_{L,n,1}$ (defined by (5.20) and (5.21)) are well-defined inner products on $\mathcal{A}^{\mathcal{S}}$ and are explicitly given by*

$$\langle f, g\rangle_{B,n,1} = \frac{1}{n!}\int_{x_1=-d}^{c}\cdots\int_{x_n=-d}^{c} f(x)g(x)\left(\prod_{i=1}^{n} w_B(x_i)\right) d_qx, \tag{5.22}$$

$$\langle f, g\rangle_{L,n,1} = \frac{1}{n!}\int_{x_1=0}^{1}\cdots\int_{x_n=0}^{1} f(x)g(x)\left(\prod_{i=1}^{n} v_L(x_i)\right) d_qx \tag{5.23}$$

for $f, g \in \mathcal{A}^{\mathcal{S}}$. The corresponding multivariable big and little q-Jacobi polynomials (using Definitions 5.1 and 5.2 for $t = 1$) can be given explicitly in terms of the one-variable big and little q-Jacobi polynomials by the formulas ($\lambda \in P^+$)

$$P_\lambda^B(x; a, b, c, d; q, 1) = |S_n^\lambda|^{-1}\sum_{w\in S_n}\left(\prod_{i=1}^{n} P_{\lambda_{w^{-1}(i)}}(x_i; a, b, c, d; q)\right), \tag{5.24}$$

where P_n ($n \in \mathbb{N}_0$) are the one-variable big q-Jacobi polynomials (Definition 2.1) and $|S_n^\lambda| := \#\{w \in S_n \mid w\lambda = \lambda\}$, and

$$P_\lambda^L(x; a, b; q, 1) = |S_n^\lambda|^{-1}\sum_{w\in S_n}\left(\prod_{i=1}^{n} p_{\lambda_{w^{-1}(i)}}(x_i; a, b; q)\right), \tag{5.25}$$

where p_n ($n \in \mathbb{N}_0$) are the one-variable little q-Jacobi polynomials (Definition 2.2).

Proof. Fix $j \in \{0, \ldots, n\}$. The set of mass points $W_B^j(q^\tau)$ (given by (5.6)) is in one-to-one correspondence with

$$V_j := \{p = (p_1, \ldots, p_n) \mid 0 \le p_1 \le \cdots \le p_j,\ 0 \le p_n \le \cdots \le p_{j+1}\} \tag{5.26}$$

by the formula

$$x^{(j)}(p; \tau) = \left(cq^{p_1}, \ldots, cq^{(j-1)\tau+p_j}, -dq^{(n-j-1)\tau+p_{j+1}}, \ldots, -dq^{p_n}\right) \in W_B^j(q^\tau). \tag{5.27}$$

We first calculate $\lim_{\tau\downarrow 0}\Delta_\tau^j(x^{(j)}(p;\tau))$ for fixed $p \in V_j$, where $\Delta_\tau^j(x)$ is given by (5.4).

Rewrite Δ_τ^j as $\Delta_\tau^j = \rho_\tau^j D_\tau^j$ with

$$\rho_\tau^j(x) := \prod_{\substack{1\le i<k\le n\\ i\le j}}\frac{x_i - x_k}{x_i - q^\tau x_k}\prod_{j<l<m\le n}\frac{x_m - x_l}{x_m - q^\tau x_l} \tag{5.28}$$

and

$$D_\tau^j(x) := \prod_{\substack{1\le i<k\le n\\ i\le j}} x_i^{2\tau}\left(q^{1-\tau}\frac{x_k}{x_i}; q\right)_{2\tau}\prod_{j<l<m\le n}|x_m|^{2\tau}\left(q^{1-\tau}\frac{x_l}{x_m}; q\right)_{2\tau}. \tag{5.29}$$

For $p \in V_j$ and $\tau \in (0, \infty)$, define

$$g_{ik}^{(j)}(p,\tau) := \frac{(x_i^{(j)}(p;\tau) - x_k^{(j)}(p;\tau))}{(x_i^{(j)}(p;\tau) - q^\tau x_k^{(j)}(p;\tau))}; \tag{5.30}$$

then we can write ρ_τ^j, evaluated at the mass point $x^{(j)}(p;\tau)$, as

$$\rho_\tau^j(x^{(j)}(p;\tau)) = \prod_{\substack{1\le i<k\le n\\ i\le j}} g_{ik}^{(j)}(p,\tau) \prod_{j<l<m\le n} g_{ml}^{(j)}(p,\tau). \tag{5.31}$$

Let $p \in V_j$; then for $1 \le i < k \le n$ with $i \le j$, we have

$$g_{ik}^{(j)}(p,0) := \lim_{\tau\downarrow 0} g_{ik}^{(j)}(p,\tau) = \begin{cases} 1 & \text{if } x_i^{(j)}(p;0) \neq x_k^{(j)}(p;0), \\ \dfrac{k-i}{k-i+1} & \text{if } x_i^{(j)}(p;0) = x_k^{(j)}(p;0), \end{cases} \tag{5.32}$$

and for $j < l < m \le n$, we have

$$g_{ml}^{(j)}(p,0) := \lim_{\tau\downarrow 0} g_{ml}^{(j)}(p,\tau) = \begin{cases} 1 & \text{if } x_l^{(j)}(p;0) \neq x_m^{(j)}(p;0), \\ \dfrac{m-l}{m-l+1} & \text{if } x_l^{(j)}(p;0) = x_m^{(j)}(p;0). \end{cases} \tag{5.33}$$

Thus we have $\lim_{\tau\downarrow 0} \rho_\tau^j(x^{(j)}(p;\tau)) = n(x^{(j)}(p;0))$ with $n(x)$ defined by

$$n(x) := \prod_{\{1\le l<m\le n \,|\, x_l = x_m\}} \frac{m-l}{m-l+1}.$$

Fix $x \in \mathbb{R}^n$ with $x_1 \ge \cdots \ge x_n$, and denote $m(x) := \#\{w \in S_n \,|\, wx = x\}$. Let $(\lambda_1, \ldots, \lambda_p)$ be the sequence of natural numbers such that $\lambda_1 + \cdots + \lambda_p = n$ and such that

$$x_1 = \cdots = x_{\lambda_1} > x_{\lambda_1+1} = \cdots = x_{\lambda_1+\lambda_2} > \cdots > x_{\lambda_1+\cdots+\lambda_{p-1}+1} = \cdots = x_n.$$

Then we have

$$\begin{aligned} n(x) = \prod_{i=1}^p \left(\prod_{1\le l<m\le\lambda_i} \frac{m-l}{m-l+1} \right) &= \prod_{i=1}^p \left(\prod_{r=1}^{\lambda_i-1} (\frac{r}{r+1})^{\lambda_i - r} \right) \\ &= \prod_{i=1}^p \frac{1}{\lambda_i!} = \frac{1}{m(x)}. \end{aligned}$$

It follows that $\lim_{\tau\downarrow 0} \rho_\tau^j(x^{(j)}(p;\tau)) = 1/m(x^{(j)}(p;0))$ for all $p \in V_j$. Furthermore, it is easily checked that $\lim_{\tau\downarrow 0} D_\tau^j(x^{(j)}(p;\tau)) = 1$ for all $p \in V_j$. Hence we have

$$\lim_{\tau\downarrow 0} \Delta_\tau^j(x^{(j)}(p;\tau)) = 1/m(x^{(j)}(p;0)), \quad p \in V_j.$$

Furthermore, we have $\lim_{\tau\downarrow 0} d_j^\tau = 1$ (with d_j^τ given by (5.5)), so for every $p \in V_j$,

$$\lim_{\tau\downarrow 0} w_j(x^{(j)}(p;\tau); q^\tau) = \frac{1}{m(x^{(j)}(p;0))} \prod_{i=1}^n w_B(x_i^{(j)}(p;0))$$

with $w_j(x)$ given by (5.3). Write $\langle f, g\rangle_{j,B,q^\tau}$ as a sum over $p \in V_j$ using formula (5.27); then we will see in Proposition 6.1 that we are allowed to pull the limit $\tau \downarrow 0$ through the infinite sum. Thus we obtain

$$\begin{aligned} \lim_{\tau\downarrow 0} \langle f, g\rangle_{B,q^\tau} = \sum_{j=0}^n \int_{x_1=0}^c \int_{x_2=0}^{x_2} \cdots \int_{x_j=0}^{x_{j-1}} \int_{x_{j+1}=-d}^0 \int_{x_{j+2}=-d}^{qx_{j+1}} \\ \cdots \int_{x_n=-d}^{qx_{n-1}} f(x)g(x) \frac{1}{m(x)} \left(\prod_{i=1}^n w_B(x_i) \right) d_q x. \end{aligned} \tag{5.34}$$

(5.22) now follows by symmetrizing the right-hand side of this formula. Formula (5.24) then follows directly from the orthogonality of the one-variable big q-Jacobi polynomials. The proof of the proposition for the little q-Jacobi case is similar. □

Note that Theorems 5.6 and 5.7 are still valid for $t = 1$. The two theorems are easy consequences of the corresponding results in the one-variable case. For $t = 1$, both theorems also follow by a continuity argument in t from the corresponding theorems for $t \in (0, 1)$.

Similarly, we can express the multivariable big (resp. little) q-Jacobi polynomials for $t = q$ in terms of one-variable big (resp. little) q-Jacobi polynomials.

Proposition 5.9. *For $\lambda \in P^+$, define $\tilde{\lambda} \in P^+$ by $\tilde{\lambda} := \lambda + \delta$ (δ given by* (4.6)*).*

(1) *Let $(a, b, c, d) \in V_B^q$ and $\lambda \in P^+$; then*

$$P_\lambda^B(x; a, b, c, d; q, q) = \Delta(x)^{-1} \sum_{w \in S_n} \det(w) \left(\prod_{i=1}^n P_{\tilde{\lambda}_{w^{-1}(i)}}(x_i; a, b, c, d; q) \right).$$

(2) *Let $(a, b) \in V_L^q$ and $\lambda \in P^+$; then*

$$P_\lambda^L(x; a, b; q, q) = \Delta(x)^{-1} \sum_{w \in S_n} \det(w) \left(\prod_{i=1}^n p_{\tilde{\lambda}_{w^{-1}(i)}}(x_i; a, b; q) \right).$$

Proof. The proposition follows from (4.7), (5.17), (5.18), (5.19), and the orthogonality in the one variable case. □

The families of multivariable big (resp. little) q-Jacobi polynomials for $t = 1$ and $t = q$ are more or less the trivial families since full orthogonality of the multivariable big (resp. little) q-Jacobi polynomials for $t = 1$ and $t = q$ can easily be deduced from the orthogonality in the one-variable case. From this point of view, we can think of t as an extra (continuous) deformation parameter linking these two trivial families of multivariable big (resp. little) q-Jacobi polynomials.

6. Some proofs. In this section, we give the proofs that we omitted in section 5. We start with the proof of Proposition 5.5.

Proposition 6.1. *Let $f, g \in \mathcal{A}^{\mathcal{S}}$.*

(1) *$\langle f, g \rangle_{B,t}$ is continuous in t for $t \in (0, 1]$, where for $t = 1$, the inner product is given by* (5.22).

(2) *$\langle f, g \rangle_{L,t}$ is continuous in t for $t \in (0, 1]$, where for $t = 1$, the inner product is given by* (5.23).

Proof. It is sufficient to prove continuity in τ for $\tau \in [0, \infty)$ ($t = q^\tau$). If $h(z, \tau)$ is a function such that $h(uq^k, \tau)$ is continuous in $\tau \in [0, \infty)$ for all $k \in \mathbb{N}_0$ ($u \neq 0$), then

$$\int_0^u h(z, \tau) d_q z = (1 - q) \sum_{k=0}^\infty h(uq^k, \tau) uq^k$$

will be continuous in τ if for every compact subset K of $[0, \infty)$, there exists a $\epsilon_K < 1$ such that

$$\sup_{(k,\tau) \in \mathbb{N}_0 \times K} \left| q^{k\epsilon_K} h(uq^k, \tau) \right| < \infty.$$

For $\tau \in [0, \infty)$, define

$$w_j(x; q^\tau) := d_j^\tau \left(\prod_{i=1}^n w_B(x_i) \right) \Delta_\tau^j(x), \tag{6.1}$$

where d_j^τ is given by (5.5) and $\Delta_\tau^j(x) := \rho_\tau^j(x) D_\tau^j(x)$ for $\tau \in [0,\infty)$, with $D_j(x)$ defined by (5.29) for $\tau \in [0,\infty)$ and with $\rho_\tau^j(x)$ defined by (5.28) for $\tau \in (0,\infty)$ and $\rho_0^j(x) := m(x)^{-1}$ for $\tau = 0$ (where $m(x) := \{w \in S_n \,|\, wx = x\}$). The inner product $\langle\, . \,, \, . \,\rangle_{B,n,q^\tau}$ for $\tau \in [0,\infty)$ can now be given by formulas (5.1) and (5.2) if we use the weight function $w_j(x;q^\tau)$ given by (6.1). (Indeed, w_j is exactly the weight function (5.3) for $\tau > 0$, and for $\tau = 0$ we have $D_0^j(x) = 1$ and $d_j^0 = 1$, so it then follows from formula (5.34).)

Therefore, in the big q-Jacobi case, it will be sufficient to prove that for every $K \subset [0,\infty)$ compact and all $j \in \{0,\ldots,n\}$,

$$\sup_{(p,\tau)\in V_j\times K} \left|\Delta_\tau^j(x^{(j)}(p;\tau))\right| < \infty, \quad j = 0,\ldots,n,$$

with V_j and $x^{(j)}(p;\tau)$ defined by (5.26) and (5.27), respectively. (Clearly, $d_j^\tau(c,d,q)$ is continuous in τ for $\tau \in [0,\infty)$.)

Similarly, in the little q-Jacobi case, we can take (5.7) as the definition of $\langle\, . \,, \, . \,\rangle_{L,n,q^\tau}$ for all $\tau \in [0,\infty)$ if we take the function $v(x;1) := \bigl(\prod_{i=1}^n v_L(x_i)\bigr)\Delta_0(x)$ with $\Delta_0(x) := \rho_0^n(x) D_0^n(x) = m(x)^{-1}$ as the weight function for $\tau = 0$ in (5.7).

Thus in the little q-Jacobi case, it will be sufficient to prove that for arbitrary $K \subset [0,\infty)$ compact,

$$\sup_{(p,\tau)\in V_n\times K} |\Delta_\tau(\tilde{x}(p;\tau))| < \infty$$

(where $\tilde{x}(p;\tau) := (q^{p_1}, q^{\tau+p_2}, \ldots, q^{(n-1)\tau+p_n})$) because if $-1 < \alpha < 0$, then the factor x_i^α in the weight function can be compensated by taking $\epsilon_{K,i} = -\alpha$ $(i = 1,\ldots,n)$.

We will prove that for all $j \in \{0,\ldots,n\}$ and all $K \subset [0,\infty)$ compact,

$$\sup_{(p,\tau)\in V_j\times K} |\rho_\tau^j(x^{(j)}(p;\tau))| < \infty \tag{6.2}$$

and

$$\sup_{(p,\tau)\in V_j\times K} |D_\tau^j(x^{(j)}(p;\tau))| < \infty. \tag{6.3}$$

Then we are ready because the little q-Jacobi case follows from (6.2) and (6.3) with $j = n$ and $c = 1$. We use the expression for $\rho_\tau^{(j)}$ evaluated at a specific mass point $x^{(j)}(p;\tau) \in W_B^j(q^\tau)$ as was given in the proof of Proposition 5.8 (formula (5.31)),

$$\rho_\tau^j(x^{(j)}(p;\tau)) = \prod_{\substack{1\le i<k\le n\\ i\le j}} g_{ik}^{(j)}(p,\tau) \prod_{j<l<m\le n} g_{ml}^{(j)}(p,\tau) \tag{6.4}$$

with $g_{rs}^{(j)}(p,\tau)$ defined by (5.30) for $\tau > 0$ and by (5.32) and (5.33) for $\tau = 0$.

Proof of (6.2). We look at the factors of the form $g_{rs}^{(j)}(l,\tau)$ in the expression for $\rho_\tau^j(x^{(j)}(l,\tau))$ (see (6.4)).

Case (1): $k \le j < m$. We have

$$g_{km}^{(j)}(l,\tau) = \frac{1 + (d/c)q^{(n-m-k+1)\tau} q^{l_m - l_k}}{1 + (d/c)q^{(n-m-k+2)\tau} q^{l_m - l_k}}$$

for $l \in V_j$ and $\tau \in K$. Thus $\sup_{(l,\tau)\in V_j\times K} |g_{km}^{(j)}(l,\tau)| < \infty$ because the map $\psi : [0,\infty)\times K \to \mathbb{R}$ given by

$$\psi(x,\tau) = \frac{1+(d/c)xq^{(n-m-k+1)\tau}}{1+(d/c)xq^{(n-m-k+2)\tau}}$$

is continuous, and $\lim_{x\to\infty}\psi(x,\tau) = q^{-\tau}$ uniformly for $\tau \in K$.

Case (2): $k < m \leq j$. Let $l \in V_j$ and $\tau \in K$. We have

$$g_{km}^{(j)}(l,\tau) = \frac{1-q^{(m-k)\tau}q^{l_m-l_k}}{1-q^{(m-k+1)\tau}q^{l_m-l_k}} \tag{6.5}$$

if $\tau > 0$ and $l_m \geq l_k$ or if $\tau = 0$ and $l_m > l_k$. Furthermore, we have $g_{km}^{(j)}(l,0) = (m-k)/(m-k+1)$ if $l_k = l_m$. We have to prove that $\sup_{(l,\tau)\in V_j\times K} |g_{km}^{(j)}(l,\tau)| < \infty$. First, consider the supremum over $V_j^0 \times K$, where V_j^0 is the subset of V_j defined by $V_j^0 := \{l \in V_j \,|\, l_k = l_m\}$. Since $g_{km}^{(j)}(l,\tau) = g_{km}(0,\tau)$ independently of $l \in V_j^0$, and since $g_{km}^{(j)}(0,\tau)$ is continuous in $\tau \in [0,\infty)$, we have $\sup_{(l,\tau)\in V_j^0\times K} |g_{km}^{(j)}(l,\tau)| < \infty$. Furthermore, $\sup_{(l,\tau)\in V_j^1\times K} |g_{km}^{(j)}(l,\tau)| < \infty$ with $V_j^1 := V_j\backslash V_j^0$ follows from the fact that the map $\psi_{km} : [0,q]\times K \to \mathbb{R}$ given by

$$\psi_{km}(x,\tau) := \frac{1-q^{(m-k)\tau}x}{1-q^{(m-k+1)\tau}x}$$

is continuous, and $[0,q]\times K$ is compact.

Case (3): $j < k < m$. Similar arguments as in case (2) gives uniform boundness of $g_{mk}^{(j)}(l,\tau)$ for $l \in V_j$ and $\tau \in K$.

Proof of (6.3). We examine the factors of the form $|x_r|^{2\tau}\left(q^{1-\tau}x_s/x_r;q\right)_{2\tau}$ in the expression for $D_j^\tau(x)$ for $x \in W_B^j(q^\tau)$ (see (5.29)).

Case (1): $k \leq j < m$. For $x \in W_B^j(q^\tau)$, we have $x_k = cq^{(k-1)\tau+l_k}$ and $x_m = -dq^{(n-m)\tau+l_m}$ for some $l_m, l_k \in \mathbb{N}_0$. Using the formula

$$q^{2k\tau}\left(q^{1-\tau-k}z;q\right)_{2\tau} = \frac{\left(q^\tau z^{-1};q\right)_k}{\left(q^{-\tau}z^{-1};q\right)_k}\left(q^{1-\tau}z;q\right)_{2\tau},$$

we get

$$x_k^{2\tau}\left(q^{1-\tau}\frac{x_m}{x_k};q\right)_{2\tau} = \left(cq^{(k-1)\tau}\right)^{2\tau}\frac{\left(q^{\tau-l_m}w_{km}^{-1};q\right)_{l_k}}{\left(q^{-\tau-l_m}w_{km}^{-1};q\right)_{l_k}}\left(q^{1-\tau+l_m}w_{km};q\right)_{2\tau}$$

with $w_{km} := (-d/c)q^{(n-m-k+1)\tau}$. Then

$$\left|\frac{\left(q^{\tau-l_m}w_{km}^{-1};q\right)_{l_k}}{\left(q^{-\tau-l_m}w_{km}^{-1};q\right)_{l_k}}\right| \leq 1$$

for all $l_k, l_m \in \mathbb{N}_0$ and $\tau \in K$ because

$$1-q^{-\tau-l_m+i}w_{km}^{-1} \geq 1-q^{\tau-l_m+i}w_{km}^{-1} > 0 \quad \forall\tau\in K,\ l_m\in\mathbb{N}_0,\ i\in\{0,\ldots,l_k-1\}.$$

Choose $N_K \in \mathbb{N}_0$ such that $2\tau \leq N_K$ for all $\tau \in K$. Then

$$\left|\left(q^{1-\tau+l_m}w_{km};q\right)_{2\tau}\right| \leq \left|\left(q^{1-\tau+l_m}w_{km};q\right)_{N_K}\right|,$$

and arguments similar to those in Case (2) of the proof of (6.2) show that

$$\sup_{(l_m,\tau)\in\mathbb{N}_0\times K}\left|\left(q^{1-\tau+l_m}w_{km};q\right)_{N_K}\right|<\infty.$$

Case (2): $k<m\le j$. For $x\in W_B^j(q^\tau)$, we have that $x_k=cq^{(k-1)\tau+l_k}$ and $x_m=cq^{(m-1)\tau+l_m}$ for some $l_k,l_m\in\mathbb{N}_0$ with $l_k\le l_m$, so

$$\left|x_k^{2\tau}\left(q^{1-\tau}\frac{x_m}{x_k};q\right)_{2\tau}\right|=\left(cq^{(k-1)\tau}\right)^{2\tau}q^{2l_k\tau}\left|\frac{\left(q^{1-\tau}\left(q^{(m-k)\tau+l_m-l_k}\right);q\right)_\infty}{\left(q^{1+\tau}\left(q^{(m-k)\tau+l_m-l_k}\right);q\right)_\infty}\right|.$$

We have

$$\left|\frac{\left(q^{1-\tau}\left(q^{(m-k)\tau+l_m-l_k}\right);q\right)_\infty}{\left(q^{1+\tau}\left(q^{(m-k)\tau+l_m-l_k}\right);q\right)_\infty}\right|\le 1$$

for $\tau\in K$ and $l_m,l_k\in\mathbb{N}_0$ with $l_k\le l_m$ because

$$1-q^{1+\tau}\left(q^{(m-k)\tau+l_m-l_k}\right)q^i\ge 1-q^{1-\tau}\left(q^{(m-k)\tau+l_m-l_k}\right)q^i>0$$

for all $\tau\in K$, all $l_m,l_k\in\mathbb{N}_0$ with $l_k\le l_m$, and all $i\in\mathbb{N}_0$.

Case (3): $j<k<m$. Similar arguments as in case (2) give a uniform boundedness of

$$\left||x_m|^{2\tau}\left(q^{1-\tau}\frac{x_k}{x_m};q\right)_{2\tau}\right|$$

for $\tau\in K, x_k\in\{-dq^{(n-k)\tau+l_k}\}_{l_k\in\mathbb{N}_0}, x_m\in\{-dq^{(n-m)\tau+l_m}\}_{l_m\in\mathbb{N}_0}$, $l_m\le l_k$. □

We have the following corollary.

COROLLARY 6.2. *Let $f,g\in\mathcal{A}^{\mathcal{S}}$.*

(1) $\langle D_{B,t}f,g\rangle_{B,t}$ *is continuous in t for $t\in(0,1]$.*

(2) $\langle D_{L,t}f,g\rangle_{L,t}$ *is continuous in t for $t\in(0,1]$.*

Proof. Let $\lambda\in P^+$. The coefficients in the expansion of the symmetric polynomial $D_{n,q,t}^{a,b,c,d}m_\lambda$ with respect to the basis of monomials $\{m_\mu\mid\mu\in P^+\}$ are continuous in $t\in(0,1]$ for arbitrary fixed $a,b,c,d\in\mathbb{C}$ because they depend polynomially on t (Proposition 4.2). Now apply Proposition 6.1. □

Define the forward and backward partial q-derivatives in the ith coordinate by

$$\left(D_q^{i,+}f\right)(x):=\frac{\left(T_{q^{-1},i}f-f\right)(x)}{(1-q)x_i}\quad\text{and}\quad\left(D_q^{i,-}f\right)(x):=\frac{\left(f-T_{q,i}f\right)(x)}{(1-q)x_i},\tag{6.6}$$

respectively. In the one-variable case, we will use the notation D_q^+ and D_q^-, respectively. $D_{n,q,t}^{a,b,c,d}$ can now be written in the following form:

$$(D_nf)(x)=\sum_{i=1}^n\left(p_i(x)\left(D_q^{i,-}f\right)(x)+q_i(x)\left(D_q^{i,+}f\right)(x)\right),$$

with

$$p_i(x;a,b,c,d;q,t):=\hat{h}(x_i;a,b,c,d;q)t^{n-1}\Delta(x)^{-1}\left(T_{t,i}\Delta\right)(x),$$

$$\hat{h}(y;a,b,c,d;q) := -q(1-q)y\left(a-\frac{c}{qy}\right)\left(b+\frac{d}{qy}\right)$$

and

$$q_i(x;c,d;q,t) := \hat{g}(x_i;c,d;q)t^{n-1}\Delta(x)^{-1}(T_{t^{-1},i}\Delta)(x),$$

$$\hat{g}(y;c,d;q) := (1-q)y\left(1-\frac{c}{y}\right)\left(1+\frac{d}{y}\right).$$

We have the following result.

Proposition 6.3. (1) *Let* $i \in \{1,\ldots,n\}$ *and* $j \in \{0,\ldots,n\}$*; then*

$$\begin{aligned}\big(T_{q^{-1},i}(p_i(\,.\,;a,b,c,d;q,t)w_j(\,.\,;a,b,c,d;q,t))\big)(x)\\ = -q_i(x;c,d;q,t)w_j(x;a,b,c,d;q,t).\end{aligned} \tag{6.7}$$

(2) *Let* $i \in \{1,\ldots,n\}$*; then*

$$\begin{aligned}\big(T_{q^{-1},i}(p_i(\,.\,;b,a,1,0;q,t)v(\,.\,;a,b;q,t))\big)(x)\\ = -q_i(x;1,0;q,t)v(x;a,b;q,t).\end{aligned} \tag{6.8}$$

Proof. In Remark 5.3, we saw that $w_j(x) = \phi_j(x)w(x)$ with w given by (5.12),

$$w(x;a,b,c,d;q,t) = \left(\prod_{j=1}^{n} w_B(x_j;a,b,c,d;q)\right)\tilde{\Delta}_\tau(x),$$

and with ϕ_j a quasi-constant function (w_B given by (2.2)). Thus for (1), it is sufficient to prove (6.7) with w_j replaced by w. For (2), it will be sufficient to prove (6.8) with v replaced by

$$\left(\prod_{j=1}^{n} v_L(x_j;a,b;q)\right)\tilde{\Delta}_\tau(x)$$

(with v_L given by (2.7)).

For every $i \in \{1,\ldots,n\}$, we have

$$\left(T_{q^{-1},i}\left(\frac{(T_{t,i}\Delta)}{\Delta}\tilde{\Delta}_\tau\right)\right)(x) = \frac{\left(T_{t^{-1},i}\Delta\right)(x)}{\Delta(x)}\tilde{\Delta}_\tau(x),$$

which follows from a straightforward calculation using (4.1). Thus the proposition follows from

$$\hat{h}(q^{-1}y;a,b,c,d;q)w_B(q^{-1}y;a,b,c,d;q) = -\hat{g}(y;c,d;q)w_B(y;a,b,c,d;q)$$

and

$$\hat{h}(q^{-1}y;b,a,1,0;q)v_L(q^{-1}y;a,b;q) = -\hat{g}(y;1,0;q)v_L(y;a,b;q). \qquad \square$$

The self-adjointness of D_B with respect to $\langle\,.\,,.\,\rangle_B$ and the self-adjointness of D_L with respect to $\langle\,.\,,.\,\rangle_L$ can now be proved with the help of the following special version of the q-partial integration rule. The proof is similar to the proof of the usual q-partial integration rule (cf. [13]).

Lemma 6.4. *Let $\alpha \neq 0$ and let f_i be functions in one variable ($i = 1, \dots, 4$). Suppose that f_1 and f_4 are defined on $\{\alpha q^k \mid k \in \mathbb{N}_0\}$ and that f_2 and f_3 are defined on $\{\alpha q^k \mid k \in \mathbb{N}_0 \cup \{-1\}\}$. Suppose that*

$$\lim_{k\to\infty} \left(f_1(\alpha q^{k+1}) f_2(\alpha q^k) + f_3(\alpha q^k) f_4(\alpha q^{k+1})\right)$$

exists and is equal to L. Then

$$\begin{aligned}\int_0^\alpha \left(\left(D_q^- f_1\right)(x) f_2(x) + \left(D_q^+ f_3\right)(x) f_4(x)\right) d_q x &= f_1(\alpha) f_2(q^{-1}\alpha) \\ + f_3(q^{-1}\alpha) f_4(\alpha) - L - \int_0^\alpha &\left(f_1(x)\left(D_q^+ f_2\right)(x) + f_3(x)\left(D_q^- f_4\right)(x)\right) d_q x.\end{aligned}$$

Theorem 6.5. *Let $t \in (0, 1]$.*

(i) *$D_{B,t}$ is self-adjoint with respect to $\langle\, .\, ,\, .\, \rangle_{B,t}$.*

(ii) *$D_{L,t}$ is self-adjoint with respect to $\langle\, .\, ,\, .\, \rangle_{L,t}$.*

Proof. (i) Fix $j \in \{0, \dots, n\}$. Define

$$W_j(\tau) := \{x \mid x_k = cq^{(k-1)\tau + l_k} (k \leq j), x_k = -dq^{(n-k)\tau + l_k} (k > j) \text{ and } l_k \in \mathbb{N}_0\}.$$

First, let us check that if $\tau \in (0, \infty) \setminus \cup_{p=1}^n (1/p)\mathbb{N}$, then $w_j(x; q^\tau) \neq 0$ for $x \in W_j(\tau)$ iff $x \in W_B^j(q^\tau)$.

Therefore, let $x \in W_j(\tau)$. Then $\left(q^{1-\tau} x_{k+1}/x_k; q\right)_\infty = 0$ if $1 \leq k < j$ and $l_k > l_{k+1}$, and $\left(q^{1-\tau} x_k/x_{k+1}; q\right)_\infty = 0$ if $j < k < n$ and $l_k < l_{k+1}$. Furthermore, if we assume that $\tau \in (0, \infty) \setminus \cup_{p=1}^n (1/p)\mathbb{N}$, then $(q^\tau x_m/x_k; q)_\infty \neq 0$ for $1 \leq k < m \leq j$ and $(q^\tau x_k/x_m; q)_\infty \neq 0$ for $j < k < m \leq n$. Therefore, if $\tau \in (0, \infty) \setminus \cup_{p=1}^n (1/p)\mathbb{N}$, then $\Delta_\tau^j(x) \neq 0$ for $x \in W_j(\tau)$ iff $x \in W_B^j(q^\tau)$, and so this also holds for $w_j(x; q^\tau)$.

As a consequence, we have that

$$\langle f, g\rangle_{j,B,t} = \int_{x_1=0}^{c} \cdots \int_{x_j=0}^{ct^{j-1}} \int_{x_{j+1}=-dt^{n-j-1}}^{0} \cdots \int_{x_n=-d}^{0} f(x) g(x) w_j(x; t) d_q x \tag{6.9}$$

for all $f, g \in \mathcal{A}^{\mathcal{S}}$ if $\tau \in (0, \infty) \setminus \cup_{p=1}^n (1/p)\mathbb{N}$. We will prove self-adjointness for $\tau \notin \cup_{p=1}^n (1/p)\mathbb{N}$; then Corollary 6.2 asserts self-adjointness for $\tau \in [0, \infty)$. Therefore, fix $\tau \in (0, \infty) \setminus \cup_{p=1}^n (1/p)\mathbb{N}$. We will apply Lemma 6.4 repeatedly on the right-hand side of the formula

$$\begin{aligned}\langle D_{B,t} f, g\rangle_{B,t} = \sum_{j=0}^{n} \sum_{l=1}^{n} \int_{x_1=0}^{c} \cdots \int_{x_j=0}^{ct^{j-1}} \int_{x_{j+1}=-dt^{n-j-1}}^{0} \\ \cdots \int_{x_n=-d}^{0} \left(p_l(x)\left(D_q^{l,+} f\right)(x) + q_l(x)\left(D_q^{l,-} f\right)(x)\right) g(x) w_j(x) d_q x.\end{aligned} \tag{6.10}$$

Formula (6.10) is valid because the proof of Proposition 6.1, together with the fact that $\tau \notin \cup_{p=1}^n (1/p)\mathbb{N}$, shows that the multisums in the right-hand side of (6.10) converge absolutely for each $j \in \{0, \dots, n\}$ and each $l \in \{1, \dots, n\}$. We are therefore also allowed to interchange the order of q-integration in the right-hand side of (6.10).

For $x = (x_1, \ldots, x_n)$, we denote $\hat{x}_i := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, and we denote $\hat{x}_i(u) := (x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_n)$. Define W_j^i by

$$W_j^i(\tau) := \{\hat{x}_i \,|\, x \in W_j(\tau)\}.$$

Note that $W_{i-1}^i = W_i^i$ for all $i \in \{1, \ldots, n\}$. Let $f, g \in \mathcal{A}^{\mathcal{S}}$. For $i \leq j$, we apply Lemma 6.4 on

$$\int_{x_i=0}^{ct^{i-1}} \big(p_i(x)\big(D_q^{i,-}f\big)(x) + q_i(x)\big(D_q^{i,+}f\big)(x)\big)g(x)w_j(x)d_qx_i, \tag{6.11}$$

and for $i > j$, we apply Lemma 6.4 to

$$\int_{x_i=0}^{-dt^{n-i}} \big(p_i(x)\big(D_q^{i,-}f\big)(x) + q_i(x)\big(D_q^{i,+}f\big)(x)\big)g(x)w_j(x)d_qx_i \tag{6.12}$$

with fixed $\hat{x}_i \in W_j^i$ for the variable x in the integrand of (6.11) and (6.12). Therefore, define

$$\begin{aligned} f_1^{i,j,\hat{x}_i}(y) &:= f(\hat{x}_i(y)), \qquad & f_2^{i,j,\hat{x}_i}(y) &:= p_i(\hat{x}_i(y))w_j(\hat{x}_i(y))g(\hat{x}_i(y)),\\ f_3^{i,j,\hat{x}_i}(y) &:= f(\hat{x}_i(y)), \qquad & f_4^{i,j,\hat{x}_i}(y) &:= q_i(\hat{x}_i(y))w_j(\hat{x}_i(y))g(\hat{x}_i(y)). \end{aligned}$$

Then formula (6.7) gives

$$\begin{aligned} &\int_{x_i=0}^{ct^{i-1}} \left(f_1^{i,j,\hat{x}_i}(x_i)\left(D_q^+ f_2^{i,j,\hat{x}_i}\right)(x_i) + f_3^{i,j,\hat{x}_i}(x_i)\left(D_q^- f_4^{i,j,\hat{x}_i}\right)(x_i)\right) d_qx_i\\ &\qquad = -\int_{x_i=0}^{ct^{i-1}} f(x)\big(p_i(x)\big(D_q^{i,-}g\big)(x) + q_i(x)\big(D_q^{i,+}g\big)(x)\big)w_j(x)d_qx_i \end{aligned} \tag{6.13}$$

if $i \leq j$ and

$$\begin{aligned} &\int_{x_i=0}^{-dt^{n-i}} \left(f_1^{i,j,\hat{x}_i}(x_i)\left(D_q^+ f_2^{i,j,\hat{x}_i}\right)(x_i) + f_3^{i,j,\hat{x}_i}(x_i)\left(D_q^- f_4^{i,j,\hat{x}_i}\right)(x_i)\right) d_qx_i\\ &\qquad = -\int_{x_i=0}^{-dt^{n-i}} f(x)\big(p_i(x)\big(D_q^{i,-}g\big)(x) + q_i(x)\big(D_q^{i,+}g\big)(x)\big)w_j(x)d_qx_i \end{aligned} \tag{6.14}$$

if $i > j$, with $\hat{x}_i \in W_j^i$ fixed for the variable x in the integrands. Define

$$h^{i,j,\hat{x}_i}(\gamma) := f_1^{i,j,\hat{x}_i}(\gamma)f_2^{i,j,\hat{x}_i}(q^{-1}\gamma) + f_3^{i,j,\hat{x}_i}(q^{-1}\gamma)f_4^{i,j,\hat{x}_i}(\gamma).$$

Then we will prove the following:

(1a) $h^{i,j,\hat{x}_i}(cq^{(i-1)\tau}) = 0$ if $i \leq j$ and $\hat{x}_i \in W_j^i$;

(1b) $h^{i,j,\hat{x}_i}(-dq^{(n-i)\tau}) = 0$ if $i > j$ and $\hat{x}_i \in W_j^i$;

(2a) $\lim_{l_i\to\infty} h^{i,j,\hat{x}_i}(cq^{(i-1)\tau+l_i+1}) = 0$ for $i \in \{1, \ldots, n\}, j \geq i+1$, and $\hat{x}_i \in W_j^i$;

(2b) $\lim_{l_i\to\infty} h^{i,j,\hat{x}_i}(-dq^{(n-i)\tau+l_i+1}) = 0$ for $i \in \{1, \ldots, n\}, j \leq i-2$, and $\hat{x}_i \in W_j^i$;

(2c) $\lim_{l_i\to\infty} h^{i,i,\hat{x}_i}(cq^{(i-1)\tau+l_i+1})$ and $\lim_{l_i\to\infty} h^{i,i-1,\hat{x}_i}(-dq^{(n-i)\tau+l_i+1})$ exist and have the same limit for all $i \in \{1, \ldots, n\}$ and all $\hat{x}_i \in W_i^i = W_{i-1}^i$.

Observe that if these five statements are valid, then the multisum over $\hat{x}_i \in W_i^i$ of the function

$$\hat{h}_i(\hat{x}_i) := \left(\prod_{j\neq i} |x_j|\right)\left(\lim_{l_i\to\infty} h^{i,i,\hat{x}_i}(cq^{(i-1)\tau+l_i+1})\right)$$

is absolutely convergent for $i = 1, \ldots, n$. This follows from the formula

$$\hat{h}_i(\hat{x}_i) := \left(\prod_{j\neq i} |x_j|\right)\left(\int_{x_i=0}^{ct^{i-1}} f(x)\big(p_i(x)\big(D_q^{i,-}g\big)(x) + q_i(x)\big(D_q^{i,+}g\big)(x)\big)w_i(x)d_qx_i\right.$$
$$\left. - \int_{x_i=0}^{ct^{i-1}} \big(p_i(x)\big(D_q^{i,-}f\big)(x) + q_i(x)\big(D_q^{i,+}f\big)(x)\big)g(x)w_i(x)d_qx_i\right),$$

which is a consequence of Lemma 6.4, (6.13), (1a), and (2c). The self-adjointness of $D_{B,t}$ with respect to $\langle\, .\, ,\, .\, \rangle_{B,t}$ then follows directly from this observation and these five statements, in view of Lemma 6.4, (6.10), (6.13), and (6.14).

For the proof of the five statements, we use the fact that

$$h^{i,j,\hat{x}_i}(\gamma) = \big(f(\hat{x}_i(\gamma))g(\hat{x}_i(q^{-1}\gamma)) - f(\hat{x}_i(q^{-1}\gamma))g(\hat{x}_i(\gamma))\big)\, p_i(\hat{x}_i(q^{-1}\gamma))w_j(\hat{x}_i(q^{-1}\gamma))$$

for $\gamma \in \{cq^{(i-1)\tau+l_i}\}_{l_i\in\mathbb{N}_0}$ if $i \leq j$ and for $\gamma \in \{-dq^{(n-i)\tau+l_i}\}_{l_i\in\mathbb{N}_0}$ if $i > j$ with fixed $\hat{x}_i \in W_j^i$. This formula is a consequence of (6.7).

(1a) If $i = 1$, then $w_j(\hat{x}_1(cq^{-1})) = 0$ for $j \geq i$ because $(qx_1/c;q)_\infty$ is zero when $x_1 = cq^{-1}$. If $1 < i \leq n$, then $w_j(\hat{x}_i(cq^{(i-1)\tau-1})) = 0$ if $j \geq i$ because $x_{i-1} = cq^{(i-2)\tau+l_{i-1}}$ for certain $l_{i-1} \in \mathbb{N}_0$, so $\big(q^{1-\tau}cq^{(i-1)\tau-1}/x_{i-1};q\big)_\infty = 0$.

(1b) This is similar to the proof of (1a).

(2a) If $i \in \{1, \ldots, n\}$ and $j \geq i+1$, then $x_{i+1} = cq^{i\tau+l_{i+1}}$ for certain $l_{i+1} \in \mathbb{N}_0$, so $\big(q^{1-\tau}x_{i+1}/cq^{(i-1)\tau+l_i};q\big)_\infty = 0$ if $l_i > l_{i+1}$, and therefore $w_j(\hat{x}_i(cq^{(i-1)\tau+l_i})) = 0$ if $l_i > l_{i+1}$.

(2b)This is similar to the proof of (2a).

(2c) f, g are polynomials, so

$$\frac{f(x)(T_{q^{-1},i}g)(x) - (T_{q^{-1},i}f)(x)g(x)}{x_i} \in \mathbb{R}[x_1, \ldots, x_n]$$

is continuous as a function of x_i in $x_i = 0$ for arbitrary fixed $\hat{x}_i$. It is therefore sufficient to prove that

$$\lim_{l_i\to\infty}\Big(-dq^{(n-i)\tau+l_i}p_i(\hat{x}_i(-dq^{(n-i)\tau+l_i}))w_{i-1}(\hat{x}_i(-dq^{(n-i)\tau+l_i}))\Big)$$

and

$$\lim_{l_i\to\infty}\Big(cq^{(i-1)\tau+l_i}p_i(\hat{x}_i(cq^{(i-1)\tau+l_i}))w_i(\hat{x}_i(cq^{(i-1)\tau+l_i}))\Big)$$

exist and that they have the same limit for all $\hat{x}_i \in W_i^i$ and all $i \in \{1, \ldots, n\}$. Fix $i \in \{1, \ldots, n\}$ and $\hat{x}_i \in W_i^i$. Since $x_ip_i(x)$ is continuous as a function of x_i in $x_i = 0$, it is sufficient to prove that

$$\lim_{l_i\to\infty} d_i^\tau\Delta_\tau^i(\hat{x}_i(cq^{(i-1)\tau+l_i})) \quad \text{and} \quad \lim_{l_i\to\infty} d_{i-1}^\tau\Delta_\tau^{i-1}(\hat{x}_i(-dq^{(n-i)\tau+l_i}))$$

exist and that they have the same limit. We have

$$\Delta_\tau^i(x) = \phi_i(x) \prod_{1\le k<i} (x_k - x_i)|x_k|^{2\tau-1} \left(q^{1-\tau}\frac{x_i}{x_k}; q\right)_{2\tau-1} \times \prod_{i<m\le n} (x_i - x_m)|x_i|^{2\tau-1} \left(q^{1-\tau}\frac{x_m}{x_i}; q\right)_{2\tau-1}$$

with

$$\phi_i(x) := \left(\prod_{\substack{1\le k<m\le n\\ k<i;\, m\neq i}} (x_k - x_m)|x_k|^{2\tau-1} \left(q^{1-\tau}\frac{x_m}{x_k}; q\right)_{2\tau-1} \right) \times \prod_{i<k<m\le n} (x_k - x_m)|x_m|^{2\tau-1} \left(q^{1-\tau}\frac{x_k}{x_m}; q\right)_{2\tau-1}$$

independent of x_i and

$$\Delta_\tau^{i-1}(x) = \phi_i(x) \prod_{1\le k<i} (x_k - x_i)|x_k|^{2\tau-1} \left(q^{1-\tau}\frac{x_i}{x_k}; q\right)_{2\tau-1} \times \prod_{i<m\le n} (x_i - x_m)|x_m|^{2\tau-1} \left(q^{1-\tau}\frac{x_i}{x_m}; q\right)_{2\tau-1}.$$

Therefore, it is sufficient to prove that

$$\lim_{l_i\to\infty} d_i^\tau \sigma_i(\hat{x}_i(cq^{(i-1)\tau+l_i})) \quad \text{and} \quad \lim_{l_i\to\infty} d_{i-1}^\tau \rho_i(\hat{x}_i(-dq^{(n-i)\tau+l_i}))$$

exist and that they have the same limit, with

$$\sigma_i(x) := \prod_{i<m\le n} (x_i - x_m)|x_i|^{2\tau-1} \left(q^{1-\tau}\frac{x_m}{x_i}; q\right)_{2\tau-1},$$

$$\rho_i(x) := \prod_{i<m\le n} (x_i - x_m)|x_m|^{2\tau-1} \left(q^{1-\tau}\frac{x_i}{x_m}; q\right)_{2\tau-1}.$$

Clearly,

$$\lim_{l_i\to\infty} \rho_i(\hat{x}_i(-dq^{(n-i)\tau+l_i})) = \prod_{i<m\le n} |x_m|^{2\tau},$$

and the formula

$$q^{k(2\tau-1)}\left(q^{1-\tau-k}z; q\right)_{2\tau-1} = \frac{\left(q^\tau z^{-1}; q\right)_k}{\left(q^{1-\tau}z^{-1}; q\right)_k}\left(q^{1-\tau}z; q\right)_{2\tau-1}$$

gives that

$$\lim_{l_i\to\infty} \sigma_i(\hat{x}_i(cq^{(i-1)\tau+l_i})) = \prod_{i<m\le n} |x_m|^{2\tau} \frac{(cq^{(i-1)\tau})^{2\tau-1}\left(q^{1-\tau}\left(x_m/cq^{(i-1)\tau}\right); q\right)_{2\tau-1}}{|x_m|^{2\tau-1}\left(q^{1-\tau}\left(cq^{(i-1)\tau}/x_m\right); q\right)_{2\tau-1}}.$$

Thus

$$\lim_{l_i\to\infty}\sigma_i(\hat{x}_i(cq^{(i-1)\tau+l_i}))=\frac{d_{i-1}^{\tau}}{d_i^{\tau}}\prod_{i<m\le n}|x_m|^{2\tau}$$

for $\hat{x}_i \in W_i^i$ because the function ψ_τ given by formula (5.15) is a quasi-constant function.

(ii) Fix $\tau\in(0,\infty)\setminus\cup_{p=1}^{n}(1/p)\mathbb{N}$ and $f,g\in\mathcal{A}^{\mathcal{S}}$; then a similar argument as in the proof of (i) gives that

$$\langle f,g\rangle_L=\int_{x_1=0}^{1}\cdots\int_{x_n=0}^{q^{(n-1)\tau}}f(x)g(x)v(x)d_qx,$$

and the order of q-integration may be changed because of absolute convergence. Fix $i\in\{1,\dots,n\}$ and fix $x_k\in\{q^{(k-1)\tau+l_k}\}_{l_k\in\mathbb{N}_0}(k\neq i)$. Define

$$\begin{aligned}f_1^i(x_i)&:=f(x), & f_2^i(x_i)&:=p_i(x;b,a,1,0;q,t)v(x;a,b;q,t)g(x),\\ f_3^i(x_i)&:=f(x), & f_4^i(x_i)&:=q_i(x;1,0;q,t)v(x;a,b;q,t)g(x)\end{aligned}$$

$(x=(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n))$. It follows from formula (6.8) and Lemma 6.4 that the proof of (ii) is complete if the following two formulas are valid:

$$f_1^i(q^{(i-1)\tau})f_2^i(q^{(i-1)\tau-1})+f_3^i(q^{(i-1)\tau-1})f_4^i(q^{(i-1)\tau})=0, \tag{6.15}$$

$$\lim_{k\to\infty}\Big(f_1^i(q^{(i-1)\tau+k+1})f_2^i(q^{(i-1)\tau+k})+f_3^i(q^{(i-1)\tau+k})f_4^i(q^{(i-1)\tau+k+1})\Big)=0. \tag{6.16}$$

For the proof, we use the formula

$$\begin{aligned}f_1^i(\gamma)f_2^i(q^{-1}\gamma)+f_3^i(q^{-1}\gamma)f_4^i(\gamma)=\big(&f(\hat{x}_i(\gamma))g(\hat{x}_i(q^{-1}\gamma))-f(\hat{x}_i(q^{-1}\gamma))g(\hat{x}_i(\gamma))\big)\\ &\times p_i(\hat{x}_i(q^{-1}\gamma);b,a,1,0;q,t)v(\hat{x}_i(q^{-1}\gamma);a,b;q,t)\end{aligned}$$

with $\gamma\in\{q^{(i-1)\tau+l_i}\}_{l_i\in\mathbb{N}_0}$. This formula is a consequence of (6.8). The proof of (6.15) is similar to the proof of (1a), and (6.16) holds because

$$\lim_{l_i\to\infty}q^{(i-1)\tau+l_i}p_i(\hat{x}_i(q^{(i-1)\tau+l_i});b,a,1,0;q,t)v(\hat{x}_i(q^{(i-1)\tau+l_i});a,b;q,t)$$

is zero for $i=1,\dots,n-1$ because $v(\hat{x}_i(q^{(i-1)\tau+l_i});a,b;q,t)=0$ for l_i sufficiently big and is also zero if $i=n$ since

$$\lim_{l_n\to\infty}v(\hat{x}_n(q^{(n-1)\tau+l_n});a,b;q,t)\left(q^{(n-1)\tau+l_n}\right)^{-\alpha}$$

exists, and

$$\lim_{l_n\to\infty}\left(q^{(n-1)\tau+l_n}\right)^{\alpha+1}p_i(\hat{x}_n(q^{(n-1)\tau+l_n});b,a,1,0;q,t)=0$$

since $\alpha>-1$. □

Note added in proof. The evaluation formula for $\langle 1,1\rangle_{L,q^k}$ which we presented in Remark 5.4 for the special parameter values $k\in\mathbb{N}$ is in fact valid for all $k\in(0,\infty)$. This follows from a modified form of Askey, Habsieger, and Kadell's formula which has recently been proved by K. Aomoto in his preprint "On elliptic product formulas for Jackson integrals associated with reduced root systems."

Acknowledgment. The author thanks Prof. Tom H. Koornwinder for many useful suggestions and helpful discussions.

REFERENCES

[1] G. E. Andrews and R. Askey, *Enumeration of partitions: The role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Aigner, ed., Reidel, Boston, 1977, pp. 3–26.

[2] G. E. Andrews and R. Askey, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux, eds., Lecture Notes in Math. 1171, Springer-Verlag, New York, 1985, pp. 36–62.

[3] R. Askey, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938–951.

[4] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54 (1985), p. 319.

[5] J. F. van Diejen, *Difference Calogero–Moser systems and finite Toda chains*, J. Math. Phys., 36 (1995), pp. 1299–1323.

[6] R. J. Evans, *Multidimensional beta and gamma integrals*, Contemp. Math., 166 (1994), pp. 341–357.

[7] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Encyclopedia of Mathematics and Its Applications 35, Cambridge University Press, Cambridge, UK, 1990.

[8] L. Habsieger, *Une q-intégrale de Selberg et Askey*, SIAM J. Math. Anal., 19 (1988), pp. 1475–1489.

[9] G. J. Heckman, *An elementary approach to the hypergeometric shift operators of Opdam*, Invent. Math., 103 (1991), pp. 341–350.

[10] K. W. J. Kadell, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969–986.

[11] T. H. Koornwinder, *Askey–Wilson polynomials for root systems of type BC*, Contemp. Math., 138 (1992), pp. 189–204.

[12] T. H. Koornwinder, *Askey–Wilson polynomials as zonal spherical functions on the SU(2) quantum group*, SIAM J. Math. Anal., 24 (1993), pp. 795–813.

[13] T. H. Koornwinder, *Compact quantum groups and q-special functions*, in Representations of Lie Groups and Quantum Groups, V. Baldoni and M. A. Picardello, eds., Pitman Research Notes in Math. 311, Longman Scientific and Technical, Harlow, UK, 1994.

[14] T. H. Koornwinder, *Jacobi functions as limit cases of q-ultraspherical polynomials*, J. Math. Anal. Appl., 148 (1990), pp. 44–54.

[15] I. G. Macdonald, *Orthogonal polynomials associated with root systems*, preprint (1987).

[16] I. G. Macdonald, *A new class of symmetric functions*, Actes 20e Séminaire Lotharingen, Publ. Inst. Rech. Math. Av., Strasbourg, France, 1988, pp. 131–171.

[17] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, Oxford University Press, London, New York, 1st ed., 1979, 2nd ed., 1995.

[18] J. V. Stokman and T. H. Koornwinder, *Limit transitions for BC type multivariable orthogonal polynomials*, Report 95-19, Mathematical Preprint Series, University of Amsterdam, Amsterdam, 1995; Canad. J. Math., to appear.

[19] L. Vretare, *Formulas for elementary spherical functions and generalized Jacobi polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 805–833.

SIAM J. MATH. ANAL.
Vol. 28, No. 2, pp. 481–498, March 1997

APPROXIMATION FROM SHIFT-INVARIANT SPACES BY INTEGRAL OPERATORS*

JUNJIANG LEI†, RONG-QING JIA‡, AND E. W. CHENEY§

Abstract. We investigate approximation from shift-invariant spaces by using certain integral operators and discuss various applications of this approximation scheme. We assume that our integral operators commute with shift operators and that their kernel functions decay at a polynomial rate. We prove that the approximation order provided by such an integral operator is m if and only if the integral operator reproduces polynomials of degree up to $m-1$, where m is a positive integer. Using this result, we characterize the approximation order provided by a finitely generated shift-invariant space whose generators decay in a polynomial rate and have stable shifts. We also review some already well-studied approximation schemes such as projection, cardinal interpolation, and quasi-interpolation by considering them as special cases of integral operators.

Key words. approximation order, shift-invariant spaces, integral operators, quasi-interpolation

AMS subject classifications. 41A35, 41A63, 65D10

PII. S0036141095279869

1. Introduction.

There are many ways to construct approximation schemes associated with shift-invariant spaces. Among them are cardinal interpolation (see, e.g., [3, 8]), quasi-interpolation (see, e.g., [4, 7, 18]), projection (see, e.g., [9, 14, 16, 19]), and convolution (see, e.g., [20]). In this paper, we unify these approximation schemes in a systematic fashion by viewing them all as special cases of the approximation scheme induced by an integral operator L of the form

$$(Lf)(x) = \int K(x,y)f(y)\,dy, \quad x \in \mathbb{R}^n, \tag{1.1}$$

where the kernel K is assumed to be a complex-valued measurable function on $\mathbb{R}^n \times \mathbb{R}^n$ and the convention $\int = \int_{\mathbb{R}^n}$ has been adopted. In particular, we characterize the approximation order provided by such an integral operator. We also give a characterization of the approximation order provided by a finitely generated shift-invariant space with stable generators. All of our results are valid for approximation in $L_p(\mathbb{R}^n)$, $1 \le p \le \infty$.

A linear subspace S of $L_p(\mathbb{R}^n)$ is called shift-invariant if $f \in S$ implies $f(\cdot - \nu) \in S$ for all $\nu \in \mathbb{Z}^n$. Since our main interest lies in approximation from shift-invariant spaces, it is natural to assume that the integral operator L commutes with all shift operators T_ν on $L_p(\mathbb{R}^n)$:

$$LT_\nu = T_\nu L, \quad \nu \in \mathbb{Z}^n.$$

* Received by the editors January 13, 1995; accepted for publication (in revised form) November 29, 1995.

http://www.siam.org/journals/sima/28-2/27986.html

† Department of Mathematics, Southern Illinois University at Carbondale, Carbondale, IL 62901-4408 (jlei@math.siu.edu).

‡ Department of Mathematics, University of Alberta, Edmonton, AB T6G 2G1, Canada (jia@xihu.math.ualberta.ca). The research of this author was supported by NSERC Canada grant OGP 121336.

§ Department of Mathematics, University of Texas at Austin, Austin, TX 78712 (cheney@math.utexas.edu). The research of this author was supported by Leicester University (Great Britain) and the SERC of Great Britain.

We use T_u to denote the translation of a function f by $u \in \mathbb{R}^n$: $(T_u f)(x) = f(x-u)$, $x \in \mathbb{R}^n$. If the translation is given by a multiinteger, we call it a shift (in accordance with current parlance). It is easily seen that the commutativity described above is equivalent to the equation

$$K(x-\nu, y) = K(x, y+\nu) \quad \text{for all } \nu \in \mathbb{Z}^n \text{ and a.e. } x, y \in \mathbb{R}^n. \tag{1.2}$$

We also assume that our integral operator decays at a polynomial rate. To be precise, the kernel function K is assumed to satisfy the following two conditions:

$$\int_{\mathbb{R}^n} |K(x,\cdot)|\, dx \in L_\infty([0,1)^n) \tag{1.3}$$

and, for some nonnegative integer m,

$$\int_{\mathbb{R}^n} (1+\|y\|)^m |K(\cdot,y)|\, dy \in L_\infty([0,1)^n), \tag{1.4}$$

where the norm $\|\cdot\|$ on $\mathbb{R}^n$ is defined by

$$\|y\| := \max\{|y_1|, \ldots, |y_n|\} \quad \text{for } y = (y_1, \ldots, y_n) \in \mathbb{R}^n.$$

Conditions (1.2)–(1.4) assure that the operator L given by (1.1) is a bounded operator on $L_p(\mathbb{R}^n)$ (see Lemma 2.2). Moreover, when conditions (1.2) and (1.4) are fulfilled, we can extend the domain of the operator L to include $\Pi_m = \Pi_m(\mathbb{R}^n)$, the linear space of polynomials of (total) degree no greater than m on $\mathbb{R}^n$. We adopt the convention that $\Pi_{-1} = \{0\}$.

For $h > 0$, let σ_h be the scaling operator defined by the equation

$$\sigma_h f := f(\cdot/h).$$

If L is a linear operator on $L_p(\mathbb{R}^n)$, then we denote by L_h the operator $\sigma_h L \sigma_{1/h}$. Given a positive integer m, we say that the integral operator L provides approximation order m if for every sufficiently smooth function f in $L_p(\mathbb{R}^n)$,

$$\|L_h f - f\|_p = \mathcal{O}(h^m) \quad \text{as } h \downarrow 0.$$

Let S be a closed shift-invariant subspace of $L_p(\mathbb{R}^n)$. Then $\sigma_h(S) = \{\sigma_h f : f \in S\}$. We say that S provides approximation order m if for every sufficiently smooth function f in $L_p(\mathbb{R}^n)$,

$$\inf_{s_h \in \sigma_h(S)} \|f - s_h\|_p = \mathcal{O}(h^m) \quad \text{as } h \downarrow 0.$$

In the next two sections, we shall show that under conditions (1.2)–(1.4), the integral operator L provides approximation order m if and only if it reproduces all polynomials in Π_{m-1}. We use this result to characterize the approximation order provided by S in terms of the Strang–Fix conditions, assuming that S is generated by finitely many functions that have stable shifts and a suitable decay.

We use the standard multiindex notation as in [1]. For instance, if α and β are multiindices, then $|\alpha|$ denotes the length of α and $\alpha \le \beta$ means that α is less than or equal to β coordinatewise. For a domain D in $\mathbb{R}^n$, we denote by $\|f\|_p(D)$ the usual L_p norm of a (complex-valued) function f on D. This is simply written as $\|f\|_p$ when $D = \mathbb{R}^n$. We use $f|_E$ to denote the restriction of the function f to a subset E of its original domain. We denote by $W_p^m(\mathbb{R}^n)$ the usual Sobolev space as in [1] and by $|f|_{m,p}$ the seminorm of a function $f \in W_p^m(\mathbb{R}^n)$. We assume that all of our functions are Lebesgue measurable. We denote by $C_c(\mathbb{R}^n)$ the space of compactly supported continuous functions on $\mathbb{R}^n$ and by $C_0(\mathbb{R}^n)$ the space of continuous functions on $\mathbb{R}^n$ that vanish at infinity (see [10, p. 126]).

2. Approximation power of integral operators. In this section, we estimate the lower bound of the approximation order provided by an integral operator that commutes with shifts. In what follows, I denotes the unit cube $[0,1)^n$ in $\mathbb{R}^n$ and p is a real number such that $1 \le p \le \infty$.

THEOREM 2.1. *Let K be a kernel function satisfying conditions* (1.2)–(1.4)*, and let L be the integral operator given in* (1.1)*. If $Lq = q$ for all $q \in \Pi_{m-1}$, then*

$$\|L_h f - f\|_p \le C|f|_{m,p} h^m, \quad f \in W_p^m(\mathbb{R}^n), \tag{2.1}$$

where C is a constant independent of p, f, and h.

This result may be viewed as a generalization of the well-known Bramble–Hilbert lemma. The proof of the theorem will be given after the next two lemmas.

LEMMA 2.2. *Let K be a kernel function satisfying conditions* (1.2)–(1.4)*, and let L be the integral operator given in* (1.1)*. Let $\|L\|_p$ be the norm of L as an operator on $L_p(\mathbb{R}^n)$. Then there exists a constant M such that $\|L\|_p \le M$ for all $1 \le p \le \infty$.*

Proof. Let $k_1 := \int |K(x,\cdot)|\,dx$. Then for $y \in I$ and $\nu \in \mathbb{Z}^n$, we deduce from (1.2) that

$$k_1(y+\nu) = \int |K(x-\nu, y)|\,dx = \int |K(x,y)|\,dx.$$

This together with (1.3) implies that $k_1 \in L_\infty(\mathbb{R}^n)$. Hence for any $f \in L_1(\mathbb{R}^n)$,

$$\|Lf\|_1 \le \int\!\!\int |K(x,y)|\,|f(y)|\,dy\,dx = \int k_1(y)\,|f(y)|\,dy \le \|k_1\|_\infty \|f\|_1.$$

Similarly, the function $k_2 := \int |K(\cdot,y)|\,dy$ lies in $L_\infty(\mathbb{R}^n)$. Thus for any $f \in L_\infty(\mathbb{R}^n)$, we have $\|Lf\|_\infty \le \|k_2\|_\infty \|f\|_\infty$. Let M be the maximum of $\|k_1\|_\infty$ and $\|k_2\|_\infty$. By the Riesz–Thorin interpolation theorem (see, e.g., [10, Theorem 6.27]), we conclude that $\|Lf\|_p \le M\|f\|_p$ for all p, $1 \le p \le \infty$. □

In the next lemma, we summarize some properties of a special case of L that will be used in the proof of Theorem 2.1 and on other occasions later. Select a function χ in $C_c^\infty(\mathbb{R}^n)$ such that $\int \chi = 1$. Let

$$K(x,y) = \chi(x-y), \quad x, y \in \mathbb{R}^n.$$

For this kernel, the integral operator defined in (1.1) is the usual convolution operator $\chi *$. It is easy to verify that this kernel function satisfies conditions (1.2)–(1.4). It follows from the previous lemma that the operator $\chi *$ is bounded on $L_p(\mathbb{R}^n)$. It is clear that $\chi * f \in C^\infty(\mathbb{R}^n)$ for every locally integrable function f. Furthermore, one can verify that the operator from $L_p(\mathbb{R}^n)$ to $\ell_p(\mathbb{Z}^n)$ given by $f \mapsto (\chi * f)|_{\mathbb{Z}^n}$ is also bounded. The dilates of this operator, $\sigma_h(\chi *)\sigma_{1/h}$ for $h > 0$, are often used as an approximation tool. The convergence of this scheme as $h \downarrow 0$ stems from the fact that $\chi * 1 = 1$, i.e., the operator $\chi *$ reproduces polynomials of degree zero. To enhance approximation power, we need to choose a kernel function in such a way that the corresponding convolution operator reproduces polynomials of a higher degree. To this end, we recall the following smoothing operator J employed in [12]:

$$(Jf)(x) = \int [f(x) - \nabla_u^m f(x)]\chi(u)\, du, \quad x \in \mathbb{R}^n,$$

where m is a positive integer and ∇_u^m is the mth difference operator defined by

$$\nabla_u^m := (1 - T_u)^m, \quad u \in \mathbb{R}^n,$$

where 1 denotes the identity operator. Simple computations show that the smoothing operator J is identical to the convolution operator $\chi_m *$, where

$$\chi_m(x) := \sum_{j=1}^{m} (-1)^{j-1} j^{-n} \binom{m}{j} \chi(x/j), \quad x \in \mathbb{R}^n. \tag{2.2}$$

It is clear that χ_m satisfies all the conditions that χ does, and, moreover, $Jq = q$ for all $q \in \Pi_{m-1}$. The preceding discussion is summarized in the following lemma.

Lemma 2.3. *The operator $J = \chi_m *$ maps locally integrable functions into infinitely differentiable functions. Moreover, there is a constant C independent of p and $f \in L_p(\mathbb{R}^n)$ such that*

(a) $\|Jf\|_p \le C\|f\|_p$,

(b) $\|(Jf)|_{\mathbb{Z}^n}\|_{\ell_p} \le C\|f\|_p$,

(c) $\|Jf - f\|_p \le C|f|_{m,p}$ *for* $f \in W_p^m(\mathbb{R}^n)$, *and*

(d) $Jf = f$ *for all* $f \in \Pi_{m-1}$.

The proof of this lemma can be found in [12].

Proof of Theorem 2.1. Let $f \in W_p^m(\mathbb{R}^n)$. It suffices to show that $\|Lf - f\|_p \le C|f|_{m,p}$ since (2.1) follows from this estimate by a change of variables. Clearly,

$$\begin{aligned} \|Lf - f\|_p &\le \|Lf - LJf\|_p + \|LJf - Jf\|_p + \|Jf - f\|_p \\ &\le \{||L||_p + 1\}||f - Jf||_p + ||LJf - Jf||_p. \end{aligned} \tag{2.3}$$

By Lemma 2.2, $M := \sup_{1 \le p \le \infty} \|L\|_p < \infty$. Moreover, part (c) of Lemma 2.3 asserts that $\|Jf - f\|_p \le C|f|_{m,p}$. Consequently,

$$\{\|L\|_p + 1\}||Jf - f||_p \le (M+1)C|f|_{m,p}.$$

It thus remains to prove that

$$\|LJf - Jf\|_p \le C|f|_{m,p}. \tag{2.4}$$

Let $g := LJf - Jf$, and define $a_x(\nu) := g(x - \nu)$ for $\nu \in \mathbb{Z}^n$. Suppose $1 \le p < \infty$. Since $g \in L_p(\mathbb{R}^n)$, with $I = [0,1)^n$ we have

$$\begin{aligned} \|LJf - Jf\|_p^p = \|g\|_p^p = \int |g(x)|^p \, dx &= \sum_{\nu \in \mathbb{Z}^n} \int_{I-\nu} |g(x)|^p \, dx \\ = \sum_\nu \int_I |g(x-\nu)|^p \, dx &= \int_I \sum_\nu |g(x-\nu)|^p \, dx = \int_I ||a_x||_{\ell_p}^p \, dx. \end{aligned}$$

If $p = \infty$, then we also have

$$\|LJf - Jf\|_\infty = \sup_{x \in I} \|a_x\|_{\ell_\infty}.$$

Now it is clear that (2.4) will be established if we can show that

$$\|a_x\|_{\ell_p} \le C|f|_{m,p}, \quad x \in I, \tag{2.5}$$

for some constant C independent of f, p, and x.

Let q_z be the $(m-1)$st Taylor polynomial of Jf about $z \in \mathbb{R}^n$, and let r_z be the corresponding remainder, $Jf - q_z$. Note that for every $z \in \mathbb{R}^n$, we have $Jf(z) = q_z(z)$. Furthermore, by the assumption on L, $Lq_z = q_z$. For each $x \in I$ and $\nu \in \mathbb{Z}^n$,

$$Jf(x-\nu) = q_{x-\nu}(x-\nu) = Lq_{x-\nu}(x-\nu).$$

We therefore have

$$\begin{aligned} a_x(\nu) = g(x-\nu) &= (LJf - Jf)(x-\nu) = (LJf - q_{x-\nu})(x-\nu) \\ &= (LJf - Lq_{x-\nu})(x-\nu) = [L(Jf - q_{x-\nu})](x-\nu) \\ &= (Lr_{x-\nu})(x-\nu) = (T_\nu L r_{x-\nu})(x) = (LT_\nu r_{x-\nu})(x). \end{aligned} \tag{2.6}$$

The last step used the fact that L commutes with shifts. Next, for $x, y \in \mathbb{R}^n$ and $\nu \in \mathbb{Z}^n$, define

$$e_{x,y}(\nu) := r_{x-\nu}(y-\nu) = (T_\nu r_{x-\nu})(y).$$

Using (2.6), we can write, for each $x \in I$ and $\nu \in \mathbb{Z}^n$,

$$\begin{aligned} a_x(\nu) = (LT_\nu r_{x-\nu})(x) &= \int K(x,y)(T_\nu r_{x-\nu})(y)\,dy \\ &= \int K(x,y) r_{x-\nu}(y-\nu)\,dy = \int K(x,y) e_{x,y}(\nu)\,dy. \end{aligned} \tag{2.7}$$

To estimate $e_{x,y}(\nu)$, use the integral form of the remainder in Taylor's theorem:

$$\begin{aligned} e_{x,y}(\nu) &= r_{x-\nu}(y-\nu) = (Jf - q_{x-\nu})(y-\nu) \\ &= \int_0^1 \sum_{|\alpha|=m} \frac{m}{\alpha!}\big(D^\alpha Jf\big)\big(x-\nu+t(y-x)\big)(1-t)^{m-1}(y-x)^\alpha dt \\ &= \int_0^1 \sum_{|\alpha|=m} \frac{m}{\alpha!}(T_{-x-t(y-x)} D^\alpha Jf)(-\nu)(1-t)^{m-1}(y-x)^\alpha\,dt. \end{aligned}$$

The operators J, D^α, and T_u commute with each other. It follows that for any $x, y \in \mathbb{R}^n$ and $\nu \in \mathbb{Z}^n$,

$$|e_{x,y}(\nu)| \le m\|y-x\|^m \int_0^1 \sum_{|\alpha|=m} |(JT_{-x-t(y-x)} D^\alpha f)(-\nu)|\,dt. \tag{2.8}$$

Since $D^\alpha f \in L_p(\mathbb{R}^n)$, we can use Lemma 2.3(b) to obtain

$$\begin{aligned} \sum_{|\alpha|=m} \|(J(T_{-x-t(y-x)} D^\alpha f))|_{\mathbb{Z}^n}\|_{\ell_p} &\le C \sum_{|\alpha|=m} \|T_{-x-t(y-x)} D^\alpha f\|_{L_p(\mathbb{R}^n)} \\ &= C \sum_{|\alpha|=m} \|D^\alpha f\|_p = C|f|_{m,p}. \end{aligned} \tag{2.9}$$

Now we combine (2.8) with (2.9) and use the generalized Minkowski inequality (see, e.g., [10, p. 186]) to write

$$\tag{2.10} \begin{aligned} \|e_{x,y}\|_{\ell_p} &\leq m\|y-x\|^m \int_0^1 \sum_{|\alpha|=m} \|(JT_{-x-t(y-x)}D^\alpha f\|_{\ell_p}\, dt \\ &\leq m||y-x||^m \int_0^1 C|f|_{m,p}\, dt = mC\,\|y-x\|^m|f|_{m,p}. \end{aligned}$$

From equation (2.7), we have

$$|a_x(\nu)| \leq \int |K(x,y)|\,|e_{x,y}(\nu)|\, dy.$$

This leads to

$$\begin{aligned} \|a_x\|_{\ell_p} &\leq \int |K(x,y)|\,\|e_{x,y}\|_{\ell_p}\, dy \\ &\leq \int |K(x,y)|mC\|x-y\|^m|f|_{m,p}\, dy. \end{aligned}$$

Note that $\|x-y\| \leq 1+\|y\|$ for $x \in I$ and $y \in \mathbb{R}^n$. Hence it follows that

$$\|a_x\|_{\ell_p} \leq mC|f|_{m,p} \int (1+\|y\|)^m |K(x,y)|\, dy.$$

Taking (1.4) into account, we obtain the desired estimate (2.5), thereby completing the proof. □

We close this section by noting that when $p=\infty$, the condition that L commutes with shifts is not required (cf. [6]).

3. Upper bound. In this section, we show that the converse of Theorem 2.1 is also true. This gives an upper bound for the approximation order and hence gives a characterization of the approximation order provided by the integral operator discussed in the preceding section.

THEOREM 3.1. *Fix $m \in \mathbb{N}$ and $p \in [1,\infty]$. Let K be a kernel function satisfying conditions* (1.2) *and* (1.4)*, and let L be the integral operator given in* (1.1)*. If*

$$\tag{3.1} \|L_h f - f\|_p(I) = o(h^{m-1}) \quad \text{for all } f \in C_c^\infty(\mathbb{R}^n), \text{ as } h \downarrow 0,$$

then $Lq = q$ for all $q \in \Pi_{m-1}$.

The proof of the theorem is based on the following lemma.

LEMMA 3.2. *Fix $m \in \mathbb{N}$. Let K be a kernel function satisfying conditions* (1.2) *and* (1.4)*, and let L be the integral operator given in* (1.1)*. If*

$$\tag{3.2} \|L_h q - q\|_1(I) = o(h^{m-1}) \quad \text{for all } q \in \Pi_{m-1}, \text{ as } h \downarrow 0,$$

then $Lq = q$ for all $q \in \Pi_{m-1}$.

Proof. Our proof is motivated by the work of Lei and Jia [17]. For $-1 \leq k \leq m-1$, we shall prove that $Lq = q$ for all monomials $q(x) = x^\alpha$ with $|\alpha| = k$. This will be done by induction on k. Since $\Pi_{-1} = \{0\}$, the statement is true for $k=-1$. Suppose $0 \leq k \leq m-1$ and $Lq = q$ for all $q \in \Pi_{k-1}$. Let $q(x) = x^\alpha$, where α is some multiindex such that $|\alpha| = k$. We wish to prove $Lq = q$.

Let $h > 0$ be such that $1/h$ is an integer. Then I is the disjoint union of the cubes $h(I - \nu)$, where ν runs over the set

$$J_h := \{(\nu_1, \dots, \nu_n) \in \mathbb{Z}^n : -1/h < \nu_j \le 0 \text{ for } j = 1, \dots, n\}.$$

Thus we have

$$\|L_h q - q\|_1(I) = \sum_{\nu \in J_h} \|L_h q - q\|_1\bigl(h(I - \nu)\bigr) = h^n \sum_{\nu \in J_h} \|L\sigma_{1/h} q - \sigma_{1/h} q\|_1(I - \nu).$$

But $\sigma_{1/h} q = h^k q$; hence it follows that

$$\|L_h q - q\|_1(I) = h^{n+k} \sum_{\nu \in J_h} \|Lq - q\|_1(I - \nu). \tag{3.3}$$

By a change of variables, we obtain

$$\|Lq - q\|_1(I - \nu) = \|T_\nu L q - T_\nu q\|_1(I) = \|L(T_\nu q) - T_\nu q\|_1(I).$$

Write $T_\nu q = q + q_\nu$, where $q_\nu \in \Pi_{k-1}$. Since $Lq_\nu = q_\nu$, it follows that

$$LT_\nu q - T_\nu q = L(q + q_\nu) - (q + q_\nu) = Lq - q.$$

This shows that

$$\|Lq - q\|_1(I - \nu) = \|Lq - q\|_1(I) \quad \text{for all } \nu \in \mathbb{Z}^n. \tag{3.4}$$

Note that the number of elements in J_h is $(1/h)^n$. Therefore, (3.3) and (3.4) together give

$$\|L_h q - q\|_1(I) = h^k \|Lq - q\|_1(I).$$

Thus $\|L_h q - q\|_1(I) = o(h^{m-1})$ implies that

$$\|Lq - q\|_1(I) = o(h^{m-1-k}).$$

Since $k \le m - 1$, we conclude that $Lq - q = 0$, i.e., $Lq = q$. □

Proof of Theorem 3.1. By the Hölder inequality, with p' being the exponent conjugate to p, we have

$$\|L_h f - f\|_1(I) \le \|L_h f - f\|_p(I) \|1\|_{p'}(I) = \|L_h f - f\|_p(I).$$

Hence (3.1) implies

$$\|L_h f - f\|_1(I) = o(h^{m-1}) \quad \text{for all } f \in C_c^\infty(\mathbb{R}^n). \tag{3.5}$$

By Lemma 3.2, the theorem will be established if we can verify (3.2). For this purpose, it suffices to consider a specific function $q(x) = x^\alpha$ for $|\alpha| \le m-1$. By the C^∞-Urysohn lemma (see [10, p. 237]), there exists a function $\chi \in C_c^\infty(\mathbb{R}^n)$ such that $0 \le \chi \le 1$, $\chi = 1$ on $[-2, 2]^n$, and χ is supported on $[-3, 3]^n$. Let $f := \chi q$. Then $f \in C_c^\infty(\mathbb{R}^n)$, $f = q$ on $[-2, 2]^n$, and $|f(x)| \le |q(x)|$ for all $x \in \mathbb{R}^n$. Hence we have

$$\|L_h q - q\|_1(I) = \|L_h q - f\|_1(I) \le \|L_h(q - f)\|_1(I) + \|L_h f - f\|_1(I).$$

In view of (3.5), we only have to show that

$$\|L_h(q-f)\|_1(I) = o(h^{m-1}). \tag{3.6}$$

A point $x' \in I$ can be written as $h(x-\nu)$, where $x \in I$, $\nu \in \mathbb{Z}^n$ and $\|\nu\| \le 1/h$. We observe that

$$\begin{aligned} L_h(q-f)(x') &= L\sigma_{1/h}(q-f)(x'/h) = L\sigma_{1/h}(q-f)(x-\nu) \\ &= T_\nu L\sigma_{1/h}(q-f)(x) = LT_\nu\sigma_{1/h}(q-f)(x). \end{aligned}$$

Note that $|q(x)-f(x)| = 0$ for $\|x\| \le 2$ and $|q(x)-f(x)| \le 2|q(x)|$ for all $x \in \mathbb{R}^n$. Hence for any $h > 0$ and $x \in \mathbb{R}^n$,

$$|(LT_\nu\sigma_{1/h})(q-f)(x)| = \Big|\int K(x,y)[(T_\nu\sigma_{1/h})(q-f)](y)\,dy\Big|$$

$$\begin{aligned} &= \Big|\int K(x,y)[\sigma_{1/h}(q-f)](y-\nu)\,dy\Big| = \Big|\int K(x,y)(q-f)(hy-h\nu)\,dy\Big| \\ &= \Big|\int_{\|hy-h\nu\|>2} K(x,y)(q-f)(hy-h\nu)\,dy\Big| \le 2\int_{\|hy-h\nu\|>2} |K(x,y)||(hy-h\nu)^\alpha|\,dy. \end{aligned}$$

However, $\|hy-h\nu\| > 2$ implies that $|(hy-h\nu)^\alpha| \le \|(hy-h\nu)\|^m = h^m\|y-\nu\|^m$. Hence it follows that

$$|(LT_\nu\sigma_{1/h})(q-f)(x)| \le 2h^m \int_{\|y-\nu\|>2/h} |K(x,y)|\|y-\nu\|^m\,dy.$$

Recall that $\|\nu\| \le 1/h$. Hence $\|y-\nu\| > 2/h$ implies

$$\|y\| \ge \|y-\nu\| - \|\nu\| > 1/h \ge \|\nu\|.$$

It follows that $\|y-\nu\| \le \|\nu\| + \|y\| \le 2\|y\|$. Therefore, for $x \in I$ and $\|\nu\| \le 1/h$, we have

$$|(LT_\nu\sigma_{1/h})(q-f)(x)| \le 2h^m \int_{\|y-\nu\|>2/h} |K(x,y)|(2\|y\|)^m\,dy \le 2^{m+1}h^m M,$$

where M is a constant such that $\int |K(x,y)|\|y\|^m\,dy \le M$ for almost every $x \in I$. The existence of such a constant is guaranteed by (1.4). Consequently,

$$|L_h(q-f)(x')| \le 2^{m+1}h^m M \quad \text{for a.e. } x' \in I,$$

from which (3.6) follows at once. This completes the proof of the theorem. □

Remarks. Theorems 2.1 and 3.1 together characterize the approximation order of an integral operator with its kernel satisfying (1.2)–(1.4) by means of the degree of polynomials the operator can reproduce. Therefore, only integer approximation orders appear in our settings. Also, Theorem 3.1 remains true if the cube I in (3.1) is replaced by any nonempty open subset of $\mathbb{R}^n$.

The following corollary will play an important role in the rest of the paper.

COROLLARY 3.4. *Let S be a shift-invariant subspace of $L_p(\mathbb{R}^n)$. Let K be a kernel function satisfying conditions* (1.2)–(1.4)*, and let L be the integral operator given in* (1.1)*. Suppose that L is a projection onto S, i.e., L maps $L_p(\mathbb{R}^n)$ to S and $Lf = f$ for all $f \in S$. Then S provides approximation order m if and only if $\Pi_{m-1} \subset S$.*

Proof. If $Lq = q$ for all $q \in \Pi_{m-1}$, then by Theorem 2.1 we have

$$\|L_h f - f\|_p \le C|f|_{m,p} h^m \quad \text{for all } f \in W_p^m(\mathbb{R}^n).$$

Consequently, since $L(\sigma_{1/h} f) \in S$,

$$\inf_{s\in S} \|\sigma_h s - f\|_p \le \|\sigma_h L \sigma_{1/h} f - f\|_p = \|L_h f - f\|_p = \mathcal{O}(h^m) \quad \text{for all } f \in W_p^m(\mathbb{R}^n).$$

Hence S provides approximation order m.

For the necessity part, assume that S provides approximation order m. Consider $f \in C_c^\infty(\mathbb{R}^n)$. Then

$$\inf_{s_h \in \sigma_h(S)} \|f - s_h\|_p = \mathcal{O}(h^m).$$

But for any $s \in S$,

$$L_h(\sigma_h s) = (\sigma_h L \sigma_{1/h})(\sigma_h s) = (\sigma_h L)(s) = \sigma_h s.$$

It follows that

$$\begin{aligned} \|f - L_h f\|_p &\le \|f - \sigma_h s\|_p + \|\sigma_h s - L_h f\|_p \\ &= \|f - \sigma_h s\|_p + \|L_h(\sigma_h s - f)\|_p \\ &\le (1 + \|L\|_p)\|f - \sigma_h s\|_p. \end{aligned}$$

Here we used $\|L_h\|_p = \|L\|_p$, proved by a change of variables. From the preceding inequality, we have

$$\|L_h f - f\|_p \le (1 + \|L\|_p) \inf_{s\in S} \|f - \sigma_h s\|_p = \mathcal{O}(h^m).$$

By Theorem 3.1, $Lq = q$ for all $q \in \Pi_{m-1}$. □

4. Stable families. In this section, we study finitely generated shift-invariant spaces whose generators decay in a polynomial rate and have stable shifts.

In order to give a precise definition of "decay" for a function, we utilize a family of spaces $\mathcal{L}_m$, which originated from [13]. For $m \in \mathbb{Z}_+$, $\mathcal{L}_m$ is defined to be the space of functions f such that

$$\operatorname*{ess\,sup}_{x\in I} \sum_{\nu\in\mathbb{Z}^n} |f(x+\nu)|(1 + \|x+\nu\|)^m < \infty.$$

If $f \in \mathcal{L}_m$, then we denote by $\|f\|_{\mathcal{L}_m}$ the essential supremum in this definition. It is clear that $\mathcal{L}_m \subset L_p(\mathbb{R}^n)$ for all $1 \le p \le \infty$ and for all $m = 0, 1, 2, \ldots$. The space $\mathcal{L}_m$ is closely related to a certain Banach algebra, which we now define. Let $\mathcal{B}_m$ be the set of functions τ of the form

$$\tau(z) = \sum_{\nu\in\mathbb{Z}^n} a(\nu) z^\nu, \quad z \in \mathbb{R}^n,$$

with

$$\|\tau\|_{\mathcal{B}_m} := \sum_{\nu\in\mathbb{Z}^n} |a(\nu)|(1 + \|\nu\|)^m < \infty. \tag{4.1}$$

Here $\mathbb{R}^n$ denotes the torus in $\mathbb{C}^n$:

$$\mathbb{R}^n = \{(z_1, z_2, \dots, z_n) \ : \ z_j \in \mathbb{C} \text{ and } |z_j| = 1 \text{ for } \ j = 1, 2, \dots, n\}.$$

With the norm $\| \cdot \|_{\mathcal{B}_m}$ and the usual pointwise operations (addition and multiplication), $\mathcal{B}_m$ becomes a Banach algebra. The relationship between $\mathcal{L}_m$ and $\mathcal{B}_m$ is revealed in the next lemma. For f and g in $\mathcal{L} := \mathcal{L}_0$, define

$$[f, g](z) := \sum_{\nu \in \mathbb{Z}^n} \langle T_\nu f, g \rangle z^\nu, \quad z \in \mathbb{R}^n.$$

This series converges uniformly and absolutely. (See part (a) of the following lemma.) Here $\langle f, g \rangle := \int f(x)\bar{g}(x)\,dx$ denotes the inner product of two functions f and g on $\mathbb{R}^n$, and $\bar{g}$ means the complex conjugate of g. In the following, we use the symbol $*'$ to denote a "semidiscrete" convolution. Thus $(f*'a)(x) = \sum_{\nu \in \mathbb{Z}^n} f(x - \nu)a(\nu)$, or $f*'a = \sum_{\nu \in \mathbb{Z}^n} a(\nu) T_\nu f$.

Lemma 4.1. *Let $f, g \in \mathcal{L}_m$ and $\tau \in \mathcal{B}_m$. Write $\tau(z) = \sum_{\nu \in \mathbb{Z}^n} a(\nu) z^\nu$. Then*

(a) *$[f, g] \in \mathcal{B}_m$ and $\|[f, g]\|_{\mathcal{B}_m} \le \|f\|_{\mathcal{L}_m} \|g\|_{\mathcal{L}_m}$,*

(b) *$f*'a \in \mathcal{L}_m$ and $\|f*'a\|_{\mathcal{L}_m} \le \|f\|_{\mathcal{L}_m} \|\tau\|_{\mathcal{B}_m}$, and*

(c) *$1/\tau \in \mathcal{B}_m$ if $\tau(z) \ne 0$ for all $z \in \mathbb{R}^n$ (Wiener's lemma).*

Proof. First note that for any $v, w \in \mathbb{R}^n$,

$$1 + \|v\| \le 1 + \|v + w\| + \|w\| \le (1 + \|v + w\|)(1 + \|w\|).$$

Consequently,

$$(1 + \|v\|)^m \le (1 + \|v + w\|)^m \, (1 + \|w\|)^m. \tag{4.2}$$

Now, to show that $[f, g] \in \mathcal{B}_m$, we calculate, with the aid of inequality (4.2),

$$\begin{aligned} \sum_{\nu \in \mathbb{Z}^n} |\langle T_\nu f, g \rangle|(1 + \|\nu\|)^m &= \sum_{\nu \in \mathbb{Z}^n} \left| \int f(x - \nu)\bar{g}(x)\,dx \right| (1 + \|\nu\|)^m \\ &\le \sum_\nu \sum_\mu \int_I \big| f(x + \mu - \nu) g(x + \mu) \big| (1 + \|\nu\|)^m \,dx \\ &\le \int_I \sum_\mu \sum_\nu \big| f(x + \mu - \nu) \big| \big| g(x + \mu) \big| (1 + \|x + \mu - \nu\|)^m (1 + \|x + \mu\|)^m \,dx \\ &\le \int_I \|f\|_{\mathcal{L}_m} \|g\|_{\mathcal{L}_m} \,dx = \|f\|_{\mathcal{L}_m} \|g\|_{\mathcal{L}_m}. \end{aligned}$$

This proves part (a) of the lemma.

For part (b), we have a similar calculation:

$$\begin{aligned} &\sum_{\nu \in \mathbb{Z}^n} |(f*'a)(x + \nu)|(1 + \|x + \nu\|)^m \\ &\qquad \le \sum_\nu \sum_\mu |f(x + \nu - \mu) a(\mu)| (1 + \|x + \nu\|)^m \\ &\qquad \le \sum_\mu \sum_\nu |f(x + \nu - \mu)| |a(\mu)| (1 + \|x + \nu - \mu\|)^m (1 + \|\mu\|)^m \\ &\qquad \le \|f\|_{\mathcal{L}_m} \|\tau\|_{\mathcal{B}_m}. \end{aligned}$$

Taking the essential supremum for $x \in I$, we obtain the assertion in (b). The assertion in (c) follows immediately from Lemmas 3.2 and 3.3 of [16]. □

We now consider stability of the shifts of a finite number of functions. Let $p \in [1, \infty]$. A finite subset $\Phi = \{\phi_j\}_{j=1}^N$ of $\mathcal{L}$ is said to have L_p-stable shifts, if there are constants $A > 0$ and $B > 0$ such that for any finitely supported sequences a_j $(j = 1, \ldots, N)$,

$$A \sum_{j=1}^{N} \|a_j\|_{\ell_p} \leq \Big\| \sum_{j=1}^{N} \phi_j *' a_j \Big\|_{L_p} \leq B \sum_{j=1}^{N} \|a_j\|_{\ell_p}. \tag{4.3}$$

The above definition is equivalent to the assertion that (4.3) holds for any sequences $a_j \in \ell_p$ when $1 \leq p < \infty$ and for any $a_j \in c_0$ when $p = \infty$. (The space c_0 consists of all sequences vanishing at infinity.) Indeed, the subspace of finitely supported sequences is dense in ℓ_p $(1 \leq p < \infty)$ and dense in c_0. Furthermore, the series involved in $f *' a$ converges absolutely for any $f \in \mathcal{L}$ and $a \in \ell_p$ (cf. [13]).

Let us recall from [13] that Φ has L_p-stable shifts if and only if there are sequences $b_j \in \ell_1(\mathbb{Z}^n)$, $j = 1, \ldots, N$, such that the functions $\widetilde{\phi}_j := \phi_j *' b_j$ are dual to the functions ϕ_j in the sense that

$$\langle T_\nu \phi_j, T_\mu \widetilde{\phi}_k \rangle = \delta_{\nu\mu} \delta_{jk}, \quad j, k = 1, \ldots, N, \ \mu, \nu \in \mathbb{Z}^n, \tag{4.4}$$

where δ is the Kronecker symbol. Therefore, we may drop the affiliation L_p- from the word "stability".

Theorem 4.2. *Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$ be a subset of $\mathcal{L}$ that has stable shifts. Then $\Phi \subset \mathcal{L}_m$ if and only if $\{\widetilde{\phi}_1, \widetilde{\phi}_2, \ldots, \widetilde{\phi}_N\} \subset \mathcal{L}_m$.*

Proof. Because of the dual relationship, we need to prove only one of the two implications. Suppose that $\Phi \subset \mathcal{L}_m$. Let $\Phi(z)$ be the $N \times N$ Gramian matrix of Φ:

$$\Phi(z) := \big([\phi_j, \phi_k](z) \big)_{1 \leq j,k \leq N}, \quad z \in \mathbb{R}^n.$$

By Lemma 4.1(a), every entry of $\Phi(z)$ lies in $\mathcal{B}_m$. Since $\mathcal{B}_m$ is a Banach algebra, every minor of $\Phi(z)$, including the determinant $\det \Phi(z)$, is in $\mathcal{B}_m$. By Theorems 4.1 and 4.2 of [13], the stability of Φ implies that $\det \Phi(z) \neq 0$ for all $z \in \mathbb{R}^n$. It follows from Lemma 4.1(c) that the function $1/\det \Phi(z)$, $z \in \mathbb{R}^n$, is in $\mathcal{B}_m$. We conclude that every entry $\tau_{jk}(z)$ of the inverse matrix of the Gramian $\Phi(z)$ is in $\mathcal{B}_m$. By [13, Theorem 4.1], the dual functions of ϕ_j are given by

$$\widetilde{\phi}_j := \sum_{k=1}^{N} \phi_k *' b_{jk}, \quad j = 1, \ldots, N,$$

where b_{jk} are the sequences representing the functions τ_{jk}:

$$\tau_{jk}(z) = \sum_{\nu \in \mathbb{Z}^n} b_{jk}(\nu) z^\nu, \quad z \in \mathbb{R}^n, \ j, k = 1, \ldots, N.$$

In light of Lemma 4.1(b), $\widetilde{\phi}_j \in \mathcal{L}_m$ for $j = 1, \ldots, N$. □

5. Approximation by projection. In this section, we study the approximation order provided by projections onto some shift-invariant spaces. We employ the spaces $\mathcal{L}_m$ introduced in the preceding section.

For a finite subset $\Phi = \{\phi_j\}_{j=1}^N$ of $L_p(\mathbb{R}^n)$, we denote by $S(\Phi)_p$ the $L_p(\mathbb{R}^n)$-closure of the linear span of the functions ϕ_j $(j = 1, \ldots, N)$ and their shifts. When the generators ϕ_j have stable shifts and $p = 2$, we can use their dual functions to construct an orthogonal projection onto $S(\Phi)_2$. Since the functions in Φ and their dual functions decay sufficiently fast, we can extend the orthogonal projections to operators on $L_p(\mathbb{R}^n)$ for $1 \le p \le \infty$. These facts are formalized as follows.

Lemma 5.1. *Let $\{\phi_1, \phi_2, \ldots, \phi_N\}$ be a subset of $\mathcal{L}_m$ with stable shifts and let $\widetilde{\phi}_j$ be the dual functions* (*in the sense of* (4.4)). *Set*

$$Pf := \sum_{j=1}^{N} \sum_{\nu \in \mathbb{Z}^n} \langle f, T_\nu \widetilde{\phi}_j \rangle \, T_\nu \phi_j, \quad f \in L_p(\mathbb{R}^n),$$

$$K(x,y) := \sum_{j=1}^{N} \sum_{\nu \in \mathbb{Z}^n} \phi_j(x-\nu) \overline{\widetilde{\phi}_j}(y-\nu), \quad x, y \in \mathbb{R}^n.$$

Then the following statements are true:

(a) *P is a projection of $L_p(\mathbb{R}^n)$ onto $S(\Phi)_p$.*

(b) *For every $f \in L_p(\mathbb{R}^n)$, $(Pf)(x) = \int K(x,y) f(y)\,dy$, $x \in \mathbb{R}^n$.*

(c) *The kernel K satisfies conditions* (1.2)–(1.4).

Proof. Let S be the linear space of all functions of the form $\sum_{j=1}^{N} \phi_j *' a_j$, where a_j lies in $\ell_p(\mathbb{Z}^n)$ for $1 \le p < \infty$ and lies in $c_0(\mathbb{Z}^n)$ for $p = \infty$. Because of the stability condition in (4.3), the space S is a closed shift-invariant subspace of $L_p(\mathbb{R}^n)$. This is the smallest closed shift-invariant space containing the functions ϕ_j ($j = 1, \ldots, N$) and their shifts. Therefore, $S = S(\Phi)_p$. In other words, any function $f \in S(\Phi)_p$ has a representation of the form $\sum_{j=1}^{N} \phi_j *' a_j$. Now we see that (a) is true by virtue of the duality (4.4). The assertion in (b) comes from a straightforward calculation. It remains to verify that the kernel function K satisfies conditions (1.2)–(1.4). Condition (1.2) is obviously satisfied. From the decay properties of ϕ_j and $\widetilde{\phi}_j$, (1.3) follows easily. Indeed,

$$\operatorname*{ess\,sup}_{y \in I} \int |K(x,y)|\,dx \le \sum_{j=1}^{N} \|\phi_j\|_{L_1} \|\widetilde{\phi}_j\|_{\mathcal{L}}.$$

To verify (1.4), we use (4.2) to obtain

$$(1 + \|y\|)^m |K(x,y)| \le \sum_{j=1}^{N} \sum_{\nu \in \mathbb{Z}^n} (1 + \|y - \nu\|)^m \, |\widetilde{\phi}_j(y-\nu)| \, (1 + \|\nu\|)^m \, |\phi_j(x - \nu)|.$$

When $x \in I$, $\|\nu\| \le 1 + \|x - \nu\|$ for every $\nu \in \mathbb{Z}^n$. Consequently,

$$\begin{aligned} &\operatorname*{ess\,sup}_{x \in I} \int (1 + \|y\|)^m |K(x,y)|\,dy \\ &\qquad \le \operatorname*{ess\,sup}_{x \in I} \sum_{j=1}^{N} \sum_{\nu \in \mathbb{Z}^n} (2 + \|x - \nu\|)^m |\phi_j(x-\nu)| \int (1 + \|y\|)^m |\widetilde{\phi}_j(y)|\,dy. \end{aligned}$$

The right-hand side is finite because of the decay properties of ϕ_j and $\widetilde{\phi}_j$. □

We are now in a position to consider the approximation power of the projection operator P, which is determined by the so-called Strang–Fix conditions. Let $m \in \mathbb{N}$, $1 \le p \le \infty$, and $\Phi = \{\phi_j\}_{j=1}^{N} \subset \mathcal{L}_m$. Note that the functions ϕ_j decay fast enough so that the partial derivatives $D^\alpha \widehat{\phi}_j$ exist and are continuous for $j = 1, \ldots, N$ and

$|\alpha| \le m-1$. (Here $\widehat{\phi}$ denotes the Fourier transform of ϕ.) We say that Φ satisfies the Strang–Fix conditions of order m if there is a finite linear combination ϕ of the functions ϕ_j and their shifts such that

$$\widehat{\phi}(0) \neq 0, \tag{5.1}$$

$$D^\alpha \widehat{\phi}(2\pi\nu) = 0, \quad |\alpha| \le m-1, \quad 0 \neq \nu \in \mathbb{Z}^n. \tag{5.2}$$

THEOREM 5.2. *Let Φ be a finite subset of $\mathcal{L}_m$ that has stable shifts. Let P be the projection given in Lemma* 5.1. *Let $h > 0$ and $P_h = \sigma_h P \sigma_{1/h}$. Then*

(a) *P_h is a projection onto $\sigma_h(S(\Phi)_p)$;*

(b) *if Φ satisfies the Strang–Fix conditions of order m, then*

$$\|P_h f - f\|_p \le C|f|_{m,p} h^m, \quad f \in W_p^m(\mathbb{R}^n), \tag{5.3}$$

where C is a constant independent of f, p, and h.

Proof. To prove (a), let $s_h \in S_h$. Then $s_h = \sigma_h s$ for some $s \in S(\Phi)_p$. It follows that

$$P_h s_h = (\sigma_h P \sigma_{1/h})(\sigma_h s) = \sigma_h P s = \sigma_h s = s_h.$$

To prove (b), suppose that Φ satisfies the Strang–Fix conditions of order m. Let $q \in \Pi_{m-1}$ and write $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$. As is well known (see, e.g., [18]), there exist sequences a_j, $j = 1, \ldots, N$, such that

(i) $|a_j(\nu)| = \mathcal{O}(\|\nu\|^{m-1})$ as $\|\nu\| \to \infty$;

(ii) $q = \sum_{j=1}^N \phi_j *' a_j$.

Note that the growth property (i) and the decay properties of ϕ_j ensure that the series defining $\phi_j *' a_j$ are absolutely convergent for almost all x. By the duality relation (4.4), we have $a_j(\nu) = \langle q, T_\nu \widetilde{\phi}_j \rangle$. Thus

$$Pq = \sum_{j=1}^N \sum_{\nu \in \mathbb{Z}} a_j(\nu) T_\nu \phi_j = \sum_{j=1}^N \phi_j *' a_j = q.$$

Since $Pq = q$ for all $q \in \Pi_{m-1}$, Theorem 2.1 asserts that the estimate in (5.3) is valid. $\square$

The converse of the preceding theorem is also true. Therefore, we have the following characterization.

THEOREM 5.3. *Let Φ be a finite subset of $\mathcal{L}_m$ that has stable shifts. For $1 \le p \le \infty$, the shift-invariant space $S(\Phi)_p$ provides approximation order m if and only if Φ satisfies the Strang–Fix conditions of order m.*

Proof. The sufficiency follows from Theorem 5.2. To prove the necessity, we apply Lemma 5.1 and Corollary 3.4 to conclude that $Pq(x) = \int K(x,y)q(y)\,dy = q(x)$ for $x \in \mathbb{R}^n$ and $q \in \Pi_{m-1}$, where P and K are as given in Lemma 5.1. Let $q_\alpha(x) := x^\alpha/\alpha!$, $x \in \mathbb{R}^n$, and

$$\rho_\alpha := \sum_{j=1}^N \langle q_\alpha, \widetilde{\phi}_j \rangle \phi_j, \quad |\alpha| \le m-1.$$

Then $\rho_\alpha \in \operatorname{span}(\Phi)$. Since $q_\alpha(y) = \sum_{\beta\le\alpha} q_\beta(y-\xi)q_{\alpha-\beta}(\xi)$ for all $y, \xi \in \mathbb{R}^n$, we obtain by Lemma 5.1 that

$$\begin{aligned} q_\alpha(x) = Pq_\alpha(x) &= \int \sum_{j=1}^{N} \sum_{\nu\in\mathbb{Z}^n} \overline{\widetilde{\phi}}_j(y-\nu)\phi_j(x-\nu)q_\alpha(y)\,dy \\ &= \sum_{\beta\le\alpha} \sum_{\nu\in\mathbb{Z}^n} \sum_{j=1}^{N} \Big[\int \overline{\widetilde{\phi}}_j(y-\nu)q_\beta(y-\nu)\,dy\Big]\phi_j(x-\nu)q_{\alpha-\beta}(\nu), \end{aligned}$$

where $x \in \mathbb{R}^n$ and $|\alpha| \le m-1$. In short, we get $q_\alpha = \sum_{\beta\le\alpha} \rho_\beta *' q_{\alpha-\beta}$ for $|\alpha| \le m-1$, which implies that Φ satisfies the Strang–Fix conditions of order m. (Here and hereafter we use $f*'g$ for $f*'(g|_{\mathbb{Z}^n})$ when g is a continuous function on $\mathbb{R}^n$.) This implication is well known. For example, the argument used in [12] can be carried over verbatim to the present setting. □

A few remarks are in order. A characterization of the L_2-approximation order of a shift-invariant space is given in [2], where the conditions on decay and stability are not required. In [11], a characterization of the L_p-approximation order ($1 \le p \le \infty$) is established for the shift-invariant space generated by a compactly supported function ϕ with $\hat{\phi}(0) \neq 0$.

6. Quasi-interpolation and cardinal interpolation. In this section, we consider the approximation power provided by two important approximation schemes: quasi-interpolation and cardinal interpolation.

Let us first look at quasi-interpolation. A quasi-interpolant associated with a collection $\Phi = \{\phi_j\}_{j=1}^N \subset \mathcal{L}$ is a linear mapping $Q(\Phi, \Lambda)$, given by

$$Q(\Phi,\Lambda)f := \sum_{\nu\in\mathbb{Z}^n} \sum_{j=1}^{N} \lambda_j(T_{-\nu}f)\,T_\nu\phi_j, \tag{6.1}$$

where the λ_j's are linear functionals on $L_p(\mathbb{R}^n)$ and Λ is the collection of these λ_j's.

Recall that a bounded linear functional λ on $L_p(\mathbb{R}^n)$ $(1 \le p < \infty)$ has the representation

$$\lambda f = \int g(x)f(x)\,dx, \quad f \in L_p(\mathbb{R}^n),$$

where g is some function in $L_{p'}(\mathbb{R}^n)$ with $1/p' + 1/p = 1$. We say that λ is compactly supported if g is. Moreover, by the Riesz representation theorem (see, e.g., [10, p. 216]), a bounded linear functional λ on $C_0(\mathbb{R}^n)$ has the representation

$$\lambda f = \int f\,d\mu, \quad f \in C_0(\mathbb{R}^n),$$

where μ is a complex Borel measure on $\mathbb{R}^n$. We say that λ is compactly supported if there is a compact subset F of $\mathbb{R}^n$ such that $\mu(E \setminus F) = 0$ for every Borel subset E of $\mathbb{R}^n$. If this is the case, then the domain of λ can be extended to $C(\mathbb{R}^n)$.

The approximation order provided by a quasi-interpolant is closely related to the Strang–Fix conditions (see, e.g., $[3, 4, 5, 6, 7, 8, 18]$). This fact is substantiated again in the following theorem. Our primary interest here is to show that the results concerning approximation power provided by integral operators may be easily applied to quasi-interpolants.

Theorem 6.1. *Let $m \in \mathbb{N}$, $1 \le p \le \infty$, and $\Phi = \{\phi_j\}_{j=1}^N \subset \mathcal{L}_m$. Then there is a set of compactly supported bounded linear functionals $\Lambda = \{\lambda_j\}_{j=1}^N$ on $L_p(\mathbb{R}^n)$ for $1 \le p < \infty$ or on $C_0(\mathbb{R}^n)$ for $p = \infty$ such that the quasi-interpolant $Q(\Phi, \Lambda)$ provides approximation order m if and only if Φ satisfies the Strang–Fix conditions of order m.*

To prove this theorem we need the following.

Lemma 6.2. *Let A be any bounded linear operator on $L_p(\mathbb{R}^n)$ $(1 \le p \le \infty)$. Let J denote the operator $\chi_m *$, where χ_m is the function given in* (2.2). *The operator A provides approximation order m if and only if the composite operator AJ does so.*

Proof. By definition, $A_h = \sigma_h A \sigma_h^{-1}$, $J_h = \sigma_h J \sigma_h^{-1}$, and $(AJ)_h = \sigma_h (AJ) \sigma_h^{-1}$. Clearly, $(AJ)_h = A_h J_h$. We also note that by a change of variable, $\|A_h\|_p \le \|A\|_p$ for all $h > 0$, where $\|A_h\|_p$ and $\|A\|_p$ denote the operator norms of A_h and A on $L_p(\mathbb{R}^n)$. It follows that

$$\begin{aligned}\Big| \|A_h f - f\|_p - \|(AJ)_h f - f\|_p \Big| \\ &\le \|(A_h f - f) - (A_h J_h f - f)\|_p \\ &= \|A_h(f - J_h f)\|_p \le \|A\|_p \, \|f - J_h f\|_p.\end{aligned}$$

By Lemma 2.3(c), $\|f - J_h f\|_p = \mathcal{O}(h^m)$. The proof is complete. □

Proof of Theorem 6.1. To prove the sufficiency, we assume that Φ satisfies the Strang–Fix conditions of order m. Recall from [12] that there are finitely supported sequences a_j, $j = 1, \dots, N$, such that the function $\phi := \sum_{j=1}^N \phi_j *' a_j$ satisfies $\phi *' q = q$ for all $q \in \Pi_{m-1}$. Let Q be the operator given by

$$Qf := \phi *' (\chi_m * f), \quad f \in L_p(\mathbb{R}^n),$$

where χ_m is the function given in (2.2). By Lemma 2.3(d), we have $Qq = q$ for all $q \in \Pi_{m-1}$. Clearly, the operator Q can be written as a quasi-interpolant in the form (6.1) and as an integral operator in the form (1.1) with the kernel function

$$K(x, y) := \sum_{j=1}^N \sum_{\nu \in \mathbb{Z}^n} \chi_m *' a_j(\nu - y)\, \phi_j(x - \nu), \quad x, y \in \mathbb{R}^n.$$

This kernel function satisfies conditions (1.2)–(1.4) since a_j and χ_m are all compactly supported. By Theorem 2.1, we conclude that Q provides approximation order m.

To prove the necessity, we assume that $\Lambda = \{\lambda_j\}_{j=1}^N$ is a set of linear functionals as described in the theorem such that the quasi-interpolant $Q := Q(\Phi, \Lambda)$ provides approximation order m. Consider the composite operator $L := QJ$, where $J := \chi_m *$. By straightforward computation, we obtain

$$Lf(x) = \int K(x, y) f(y)\, dy, \quad x \in \mathbb{R}^n,$$

where

$$K(x, y) = \sum_{\nu \in \mathbb{Z}^n} \sum_{j=1}^N \lambda_j \big(\chi_m(\cdot + \nu - y)\big) \phi_j(x - \nu), \quad x, y \in \mathbb{R}^n.$$

Clearly, the kernel K satisfies (1.2). Moreover, since the λ_j's are bounded and compactly supported, K satisfies conditions (1.3) and (1.4). In light of Lemma 6.2, L

provides approximation order m. By Theorem 3.1, $Lq = q$ for all $q \in \Pi_{m-1}$. Consequently, $Qq = q$ for all $q \in \Pi_{m-1}$ by Lemma 2.3(d).

In order to complete the proof, we need to consider linear functionals on Π_{m-1}. Let Δ_{m-1} denote the set of multiindices β with $|\beta| \le m-1$. Let λ be a linear functional on Π_{m-1}. We claim that there exists a sequence a with support on Δ_{m-1} such that for all $\nu \in \mathbb{Z}^n$,

$$\lambda(q(\cdot + \nu)) = \sum_{\alpha \in \mathbb{Z}^n} a(\alpha) q(\nu + \alpha) \quad \text{for all } q \in \Pi_{m-1}. \tag{6.2}$$

Indeed, for each $\alpha \in \Delta_{m-1}$, we can find a unique polynomial $q_\alpha \in \Pi_{m-1}$ such that $q_\alpha(\alpha) = 1$ and $q_\alpha(\gamma) = 0$ for $\gamma \in \Delta_{m-1} \setminus \{\alpha\}$. Let $a(\alpha) := 0$ for $\alpha \in \mathbb{Z}^n \setminus \Delta_{m-1}$ and $a(\alpha) := \lambda q_\alpha$ for $\alpha \in \Delta_{m-1}$. Then for every $\gamma \in \Delta_{m-1}$, we have

$$\sum_{\alpha \in \mathbb{Z}^n} a(\alpha) q_\gamma(\alpha) = a(\gamma) = \lambda q_\gamma.$$

Since any polynomial $q \in \Pi_{m-1}$ can be represented as $\sum_{\gamma \in \Delta_{m-1}} q(\gamma) q_\gamma$, it follows that

$$\sum_{\alpha \in \mathbb{Z}^n} a(\alpha) q(\alpha) = \lambda q.$$

The above identity is true for all $q \in \Pi_{m-1}$. Thus if we replace q by $q(\cdot + \nu)$, where $\nu \in \mathbb{Z}^n$, we obtain the desired relation (6.2).

By what has been proved in the previous paragraph, for each $j = 1, \ldots, N$, we can find sequences a_j with support on Δ_{m-1} such that

$$\lambda_j(q(\cdot + \nu)) = \sum_{\alpha \in \mathbb{Z}^n} a_j(\alpha) q(\nu + \alpha) \quad \text{for all } q \in \Pi_{m-1}.$$

It follows that

$$Q(\Phi, \Lambda) q = \sum_{\nu \in \mathbb{Z}^n} q(\nu) \phi(\cdot - \nu),$$

where $\phi := \sum_{j=1}^N \phi_j *' a_j$. Therefore, $Q(\Phi, \Lambda) q = q$ for all $q \in \Pi_{m-1}$ implies that $\phi *' q = q$ for all $q \in \Pi_{m-1}$. Hence Φ satisfies the Strang–Fix conditions of order m. The proof of the theorem is complete. □

Now let us consider cardinal interpolation. Let ϕ be a continuous function in $\mathcal{L}$. To interpolate a bounded sequence b by the shifts of ϕ, we look for a bounded sequence a such that the function $\phi *' a$ agrees with b on $\mathbb{Z}^n$. This is often called "cardinal interpolation." The problem is said to be "poised" if for any bounded sequence b there exists a unique bounded sequence a for which $\phi *' a$ is a cardinal interpolant to b. Employing the same argument used in [3] and [8], we can show that cardinal interpolation with $\phi \in \mathcal{L}$ is poised if and only if the symbol of ϕ, defined by the equation

$$s_\phi(\xi) := \sum_{\nu \in \mathbb{Z}^n} \phi(\nu) e^{i\nu \cdot \xi}, \quad \xi \in \mathbb{R}^n,$$

does not vanish anywhere on $\mathbb{R}^n$. If this is the case, we define the function ψ by the rule

$$\widehat{\psi} := \widehat{\phi} / s_\phi. \tag{6.3}$$

The function ψ turns out to be well defined and to be the basic Lagrange interpolation function because

$$\psi(\nu) = \delta_{0\nu}, \quad \nu \in \mathbb{Z}^n.$$

Therefore, $\psi *'$ is the cardinal interpolation operator.

THEOREM 6.3. *Let $m > 0$ and let ϕ be a continuous function in $\mathcal{L}_m$. Assume that cardinal interpolation with ϕ is poised. Then the function ψ as defined in equation* (6.3) *belongs to $\mathcal{L}_m$. Furthermore, the following statements are equivalent:*

(a) *The cardinal interpolation operator $\psi *'$ provides L_∞-approximation order m.*

(b) *The function ϕ satisfies the Strang–Fix conditions of order m.*

(c) *The shift-invariant space $S(\phi)_p$ generated by ϕ in $L_p(\mathbb{R}^n)$ $(1 \le p \le \infty)$ provides L_p-approximation order m.*

Proof. First, we prove that (a) and (b) are equivalent. Let

$$Lf := \psi *'(\chi_m * f), \quad f \in L_\infty(\mathbb{R}^n),$$

where χ_m is the function given in (2.2). Then

$$Lf(x) = \int K(x,y) f(y)\, dy$$

with the kernel function

$$K(x,y) = \sum_{\nu \in \mathbb{Z}^n} \chi_m(\nu - y)\psi(x - \nu).$$

Clearly, K satisfies condition (1.2). Moreover, since $\psi \in \mathcal{L}_m$ and χ_m is compactly supported, K also satisfies conditions (1.3) and (1.4). By Lemma 6.2 and Theorems 2.1 and 3.1, we conclude that the cardinal interpolation operator $\psi *'$ provides L_∞-approximation order m if and only if $Lq = q$ for all $q \in \Pi_{m-1}$. The latter is equivalent to saying that $\psi *' q = q$ for all $q \in \Pi_{m-1}$. If this is true, then by the Poisson summation formula, we have $\hat{\psi}(0) = 1$ and $D^\alpha \hat{\psi}(2\beta\pi) = 0$ for $|\alpha| \le m-1$ and $\beta \in \mathbb{Z}^n \setminus \{0\}$. Since $\hat{\phi} = \hat{\psi} s_\phi$, by the Leibniz rule for differentiation, we see that ϕ satisfies the Strang–Fix conditions of order m. Conversely, if ϕ satisfies the Strang–Fix conditions of order m, then so does ψ because $\hat{\psi} = \hat{\phi}/s_\phi$. Hence for any $q \in \Pi_{m-1}$, there exists a sequence a such that $|a(\nu)| = \mathcal{O}(\|\nu\|^{m-1})$ as $\|\nu\| \to \infty$ and $q = \sum_{\nu \in \mathbb{Z}^n} a(\nu)\psi(\cdot - \nu)$. But $\psi(\nu) = \delta_{0\nu}$, so this gives $a(\nu) = q(\nu)$ for $\nu \in \mathbb{Z}^n$. In other words, $\psi *' q = q$ for all $q \in \Pi_{m-1}$. Thus conditions (a) and (b) are equivalent.

Second, we show that conditions (b) and (c) are equivalent. By the assumption, $s_\phi(\xi) \ne 0$ for all $\xi \in \mathbb{R}^n$. It follows that the shifts of ϕ are stable. Indeed, by the Poisson summation formula, we have

$$s_\phi(\xi) = \sum_{\beta \in \mathbb{Z}^n} \hat{\phi}(\xi + 2\pi\beta), \quad \xi \in \mathbb{R}^n.$$

Hence $s_\phi(\xi) \ne 0$ implies $\hat{\phi}(\xi + 2\pi\beta) \ne 0$ for at least one $\beta \in \mathbb{Z}^n$. Thus Theorem 5.3 is applicable to the present situation. The equivalence of (b) and (c) follows. ☐

Remark. After this work had been completed, we learned that Kyriazis in [15] also considered approximation by means of kernel operators. In his paper, Kyriazis obtained lower bounds for the approximation order provided by integral operators under a different set of conditions assumed for the kernel functions. However, his paper does not contain any result on the upper bound for the approximation order.

REFERENCES

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

[2] C. de Boor, R. DeVore, and A. Ron, *Approximation from shift-invariant subspaces of $L_2(\mathbb{R}^d)$*, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806.

[3] C. de Boor, K. Höllig, and S. Riemenschneider, *Bivariate cardinal interpolation by splines on a three-direction mesh*, Illinois J. Math., 29 (1985), pp. 533–566.

[4] C. de Boor and R. Q. Jia, *Controlled approximation and a characterization of the local approximation order*, Proc. Amer. Math. Soc., 95 (1985), pp. 547–553.

[5] H. G. Burchard and J. J. Lei, *Coordinate-wise approximation with quasi-interpolants*, J. Approx. Theory, 82 (1995), pp. 240–256.

[6] E. W. Cheney and J. J. Lei, *Quasi-interpolation on irregular points*, in Approximation and Computation: A Festschrift in Honor of Walter Gautschi, R. V. M. Zahar, ed., ISNM Series, Vol. 119, Birkhäuser Boston, Cambridge, MA, 1994, pp. 121–135.

[7] C. K. Chui and H. Diamond, *A natural formulation of quasi-interpolation by multivariate splines*, Proc. Amer. Math. Soc., 99 (1987), pp. 643–646.

[8] C. K. Chui, K. Jetter, and J. D. Ward, *Cardinal interpolation by multivariate splines*, Math. Comp., 48 (1987), pp. 711–724.

[9] A. Fischer, *Multiresolution analysis and multivariate approximation of smooth signals in $C_B(\mathbb{R}^d)$*, J. Fourier Anal. Appl., to appear.

[10] G. B. Folland, *Real Analysis*, John Wiley, New York, 1984.

[11] R. Q. Jia, *The Toeplitz theorem and its applications to approximation theory and linear partial differential equations*, Trans. Amer. Math. Soc., 347 (1995), pp. 2585–2594.

[12] R. Q. Jia and J. J. Lei, *On approximation by multi-integer translates of functions having global support*, J. Approx. Theory, 72 (1993), pp. 2–23.

[13] R. Q. Jia and C. A. Micchelli, *Using the refinement equations for the construction of pre-wavelets* II: *Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, Boston, 1991, pp. 209–246.

[14] S. E. Kelly, M. A. Kon, and L. A. Raphael, *Pointwise convergence of wavelets expansions*, Bull. Amer. Math. Soc., 30 (1994), pp. 87–94.

[15] G. C. Kyriazis, *Approximation of distribution spaces by means of kernel operators*, J. Fourier Anal. Appl., to appear.

[16] J. J. Lei, *$L_p(\mathbb{R}^d)$-Approximation by certain projection operators*, J. Math. Anal. Appl., 185 (1994), pp. 1–14.

[17] J. J. Lei and R. Q. Jia, *Approximation by piecewise exponentials*, SIAM J. Math. Anal., 22 (1991), pp. 1776–1789.

[18] W. A. Light and E. W. Cheney, *Quasi-interpolation with translates of a function having non-compact support*, Constr. Approx., 8 (1992), pp. 35–48.

[19] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[20] Y. Xu, W. A. Light, and E. W. Cheney, *Constructive methods of approximation by ridge functions and radial functions*, Numer. Algorithms, 4 (1993), pp. 205–223.

SIAM J. Math. Anal.
Vol. 28, No. 3, pp. 499–515, May 1997

GAS DYNAMICS SYSTEM: TWO SPECIAL CASES*

FRANÇOIS BEREUX[†], ERIC BONNETIER[†], AND PHILIPPE G. LeFLOCH[†]

Abstract. We consider the compressible Euler system of conservation laws for the mass, momentum, and energy of a gas. We prove that Murat and Tartar's compensated compactness method applies to this system in two cases: either

(1) the equations of state have a special form (specifically, the Lagrangian sound speed is a function of the pressure only) that leads to the existence of a large family of entropy pairs

or

(2) the conservation law of energy is replaced by the conservation law for the specific entropy. Indeed, the latter is the physically meaningful formulation when the system is derived by relaxation from an isentropic two-phase mixture.

The paper primarily investigates the structure of the gas dynamics system and includes a complete description of the (i) mathematical entropies, (ii) decoupling properties, (iii) invariant domains after Chuey, Conley, and Smoller's theory, and (iv) Tartar's commutation relations. The existence of weak solutions in cases (1) and (2) above is deduced from the pioneering work of DiPerna.

Key words. gas dynamics, discontinuous solutions, entropy, compensated compactness method

AMS subject classifications. 35L65, 65M12

PII. S0036141095285831

1. Introduction.

This paper considers certain hyperbolic systems of PDEs that arise in the modeling of compressible fluid flows and multiphase mixtures. We are interested in proving the existence of entropy discontinuous solutions based on the vanishing artificial viscosity method. We recall that smooth solutions to hyperbolic conservation laws do not exist globally in time even if the initial data are smooth. Weak solutions in the sense of distributions are sought and must be further selected with the so-called entropy criterion; cf. Lax [17, 18] and Dafermos [12] for background on hyperbolic problems.

Activity concerning the existence of entropy weak solutions to systems of conservation laws with data in L^∞ was initiated in the pioneering work of DiPerna [13, 14, 15]. In particular, in [14], the existence of globally-defined-in-time, weak solutions to the isentropic Euler system was established. This system is composed of two conservation laws for the mass and momentum of the gas assuming (formally) that the specific entropy is a constant. DiPerna's proof can be viewed as one of the main success of the compensated compactness method introduced by Murat and Tartar [22, 30]. The latter is an efficient tool for studying nonlinear composite limits of weakly convergent sequences. The proof by DiPerna uses the fact that the system of isentropic gas dynamics is genuinely nonlinear in the sense of Lax and admits a large family of mathematical entropy pairs, i.e., additional conservation laws. The latter are necessary to take advantage of Tartar's commutation relations.

DiPerna's results have been extended in several directions for systems of two (and more) conservation laws. We refer to the works by Serre [26, 28, 29] and Chen et al.

[*] Received by the editors May 10, 1995; accepted for publication February 9, 1996.
http://www.siam.org/journals/sima/28-3/28583.html

[†] Centre de Mathématiques Appliquées (CNRS, URA 756), Ecole Polytechnique, 91128 Palaiseau cedex, France (bereux@cmapx.polytechnique.fr, eopus@cmapx.polytechnique.fr, lefloch@cmapx.polytechnique.fr). The research of the third author was partially supported by the Centre National de la Recherche Scientifique (CNRS), National Science Foundation grants DMS 94-01003 and DMS 95-02766, and an NSF Faculty Early Career Development (CAREER) award.

[3, 6, 7, 8] and the references cited therein; cf. also the recent contributions by Lions et al. [19, 20] and Chen and Lefloch [5]..

We focus here on the full system of gas dynamics composed of three conservation laws of mass, momentum, and energy. Generally speaking, this system shares few properties with the isentropic system due to the lack of mathematical entropies, and it has been recognized that the compensated compactness method should fail in general [2]. In this paper, we intend to tackle this system in two *special* instances. In both cases, it will be established that the 3×3 system of gas dynamics has a very similar structure to that of the isentropic 2×2 system. Our results can therefore be viewed as a direct corollary of DiPerna's fundamental work [13, 14]. The purpose of this paper is above all to investigate the mathematical properties of the gas dynamics system. It contains a full description of the mathematical entropies, decoupling properties, bounded invariant domains, and properties of Tartar's commutation relations.

Our motivation in this paper has been to search for (necessary and) sufficient conditions that allow one to use the compensated compactness technique on the gas dynamics system. In section 2, we investigate a *special class of equations of state*—specifically, the case where the Lagrangian sound speed of the gas depends upon the pressure only. This assumption leads to many interesting properties for the Euler system. It admits a large family of mathematical entropy pairs, and Chuey, Conley, and Smoller's theory applies to obtain bounded invariant domains. It follows that under the assumption above, the system belongs to the class of rich systems introduced by Serre [27]. We prove the strong convergence of the vanishing artificial viscosity method, which implies existence of entropy weak solutions. We recall that the 3×3 system under consideration possesses two genuinely nonlinear characteristic fields and one linearly degenerate field. Observe that in our main theorem in section 2 (Theorem 2.1), all of the variables (say, density, velocity, and entropy) are proven to be L^∞ functions; furthermore, the entropy variable can propagate oscillations as expected. Note that Theorem 2.1 allows vacuum states in the solution. The system is not strictly hyperbolic at those points, and we use [14].

As this paper was completed, we learned of a result similar to our Theorem 2.1—however, in the Lagrangian framework—by Chen and Dafermos [4].

Section 3 of the paper considers a coupled set of two isentropic Euler systems that arise in the modeling of two-phase mixtures. By relaxation from this 4×4 model, a system of three conservation laws is derived. We observe that, somewhat surprisingly, and for smooth solutions only, the latter is equivalent to the gas dynamics system. However, the systems are distinct for discontinuous solutions. Instead of the classical formulation that is composed of the conservation laws of mass, momentum, and energy supplemented with the entropy inequality, we arrive at a set of conservation laws for the mass, momentum, and *specific entropy* supplemented with an *energy* inequality. In this new form, the gas dynamics system happens to share many properties with the isentropic 2×2 system. We check that the linearly degenerate characteristic field is a Temple field so that the method of Benzoni-Gavage and Serre [1] applies. We prove the existence of entropy solutions for this rather nonclassical formulation of the gas dynamics system. The key of the proof is that a uniform estimate for the total variation of the specific entropy can be derived.

For further properties of the compressible Euler equations, we refer to Bonnetier and LeFloch [2].

2. A class of equations of state.

We consider the compressible Euler system, which consists of the conservation laws for the mass, momentum, and total energy of a gas [10]:

$$\begin{aligned}\partial_t\rho + \partial_x(\rho u) &= 0,\\ \partial_t(\rho u) + \partial_x(\rho u^2 + p) &= 0,\\ \partial_t(\rho E) + \partial_x(\rho u E + pu) &= 0.\end{aligned} \tag{2.1}$$

The main unknowns are the specific density ρ, the velocity u, and the total energy E. We define the internal energy e as $e = E - u^2/2$ and the specific entropy S and the temperature $T > 0$ by the classical relation $TdS = pd(1/\rho) + de$. The equation of state yields the pressure p as a (smooth enough) function of ρ and S:

$$p = p(\rho, S) > 0, \qquad p_\rho > 0 \quad \text{for all } \rho > 0. \tag{2.2a}$$

We also make the standard assumption that

$$p_{\rho\rho} > 0 \quad \text{and} \quad p_S > 0 \quad \text{for all } \rho > 0. \tag{2.2b}$$

The Eulerian sound speed c is classically defined by

$$c^2 = p_\rho(\rho, S) = \big(p_\rho + \rho^{-2} p p_e\big)(\rho, e).$$

It will be convenient to express our results in term of the variables u, p, and S. Therefore, we set $\rho = \tilde\rho(p, S)$ and $e = \tilde e(p, S)$ with the natural constraint $\tilde e_p + p\big(1/\tilde\rho\big)_p = 0$. We introduce the specific volume $v = 1/\rho$ (we shall use the notation $v = \tilde v(p, S)$ as well) and the Lagrangian sound speed $C = \rho c$. The latter is the relevant wave speed when the mass Lagrangian coordinates are being used.

In view of (2.2), system (2.1) is strictly hyperbolic and possesses two genuinely nonlinear characteristic fields associated with the wave speeds $u \pm c$ and one linearly degenerate field associated with u. Strict hyperbolicity, however, may be lost in the vacuum where $\rho = 0$. It is well known that discontinuities form in finite time in initially smooth solutions to (2.1). We are interested in proving existence of weak solutions based on the vanishing artificial viscosity method. Therefore, we consider the following approximation:

$$\begin{aligned}\partial_t\rho^\varepsilon + \partial_x(\rho^\varepsilon u^\varepsilon) &= \varepsilon\, \partial_{xx}^2(\rho^\varepsilon),\\ \partial_t(\rho^\varepsilon u^\varepsilon) + \partial_x(\rho^\varepsilon (u^\varepsilon)^2 + p^\varepsilon) &= \varepsilon\, \partial_{xx}^2(\rho^\varepsilon u^\varepsilon),\\ \partial_t(\rho^\varepsilon E^\varepsilon) + \partial_x(\rho u^\varepsilon E^\varepsilon + p^\varepsilon u^\varepsilon) &= \varepsilon\, \partial_{xx}^2(\rho^\varepsilon E^\varepsilon).\end{aligned} \tag{2.3}$$

We shall assume that smooth solutions to (2.3) assuming a smooth initial data exist. For a proof in the isentropic case, see [14]. Initial data for u, p, and S—say, u_0, p_0, and S_0, respectively—are assumed to be given functions in $L^\infty(\mathbb{R})$ and, with obvious notation, we assume that ρ_0 and $\rho_0 E_0$ belong to $L^1(\mathbb{R})$. The entropy inequality is classically stated as

$$\partial_t(\rho S) + \partial_x(\rho u S) \le 0, \tag{2.4}$$

which, at least formally, can be checked for the limit of the $(u^\varepsilon, p^\varepsilon, S^\varepsilon)$'s as $\varepsilon \to 0$.

In this section, we consider the case where the Lagrangian sound speed is a function of the pressure, i.e.,

$$C(p, S) = C_0(p) \quad \text{or, equivalently,} \quad \rho c(p, S) = C_0(p). \tag{2.5}$$

As we prove below, (2.5) is satisfied if for some positive constants a and b,

$$\frac{1}{\tilde{\rho}(p,S)} - aS \quad \text{and} \quad \tilde{e}(p,S) - bS \quad \text{are functions of } p \text{ only.} \tag{2.6}$$

Our main result of existence and convergence is now stated.

THEOREM 2.1. *Consider the compressible Euler system* (2.1)–(2.2) *with an equation of state of the form* (2.6)*. Then we have the following uniform bounds for the approximate solutions* $(u^\varepsilon, p^\varepsilon, S^\varepsilon)$ *generated by* (2.3):

$$\begin{aligned} &0 \le p^\varepsilon(t,x), |u^\varepsilon(t,x)| \le \text{const.}\big(\|p_0\|_{L^\infty(\mathbb{R})} + \|u_0\|_{L^\infty(\mathbb{R})}\big), \\ &0 \le S^\varepsilon(t,x) \le \text{const.} \sup\nolimits_{\mathbb{R}} S_0, \end{aligned} \tag{2.7a}$$

where the constants are independent of t, x, *and* ε, *and*

$$\|\rho^\varepsilon(t)\|_{L^1(\mathbb{R})} = \|\rho_0^\varepsilon\|_{L^1(\mathbb{R})}, \qquad \|\rho^\varepsilon(t)E^\varepsilon(t)\|_{L^1(\mathbb{R})} = \|\rho_0^\varepsilon E_0\|_{L^1(\mathbb{R})}. \tag{2.7b}$$

Let us make one of the following two assumptions:

(1) *the approximations are bounded away from the vacuum, i.e.,*

$$0 < \text{const.} \le p^\varepsilon(t,x) \le \text{const.} \tag{2.8a}$$

with uniform constants;

(2) *the equation of state has the form*

$$p = k\,\rho^\gamma \quad \textit{with } \gamma > 1 \quad \textit{and} \quad k > 0. \tag{2.8b}$$

Then there is a triplet (u,p,S) *in* $L^\infty(\mathbb{R}_+ \times \mathbb{R})^3$ *with* ρ *and* ρE *in* $L^\infty(\mathbb{R}_+, L^1(\mathbb{R}))$ *and a subsequence of* $(u^\varepsilon, p^\varepsilon, S^\varepsilon)$ *such that*

$$\begin{aligned} &u^\varepsilon \to u, \qquad p^\varepsilon \to p \quad \textit{in the strong } L^q \textit{ norm for all finite } q, \\ &\rho^\varepsilon = \tilde{\rho}(p^\varepsilon, S^\varepsilon) \rightharpoonup \rho = \tilde{\rho}(p,S) \quad \textit{in the weak-}\star\ L^\infty \textit{ topology,} \end{aligned} \tag{2.9}$$

and (u,p,S) *is an entropy weak solution to* (2.1) *and* (2.4)*.*

The proof of Theorem 2.1 is based on four lemmas dealing, respectively, with

1. the mathematical entropies (Lemma 2.1),
2. a decoupling property (Lemma 2.2),
3. Tartar's equation (Lemma 2.3), and
4. the invariant regions (Lemma 2.4).

We shall state these four lemmas, give their proofs, and conclude with a proof of Theorem 2.1. First, we provide a complete description of the set of entropies. A pair (U,F) is a mathematical entropy for the system

$$\partial_t w + \partial_x f(w) = 0$$

if every smooth solution w satisfies the additional conservation law

$$\partial_t U(w) + \partial_x F(w) = 0.$$

Equivalently, the pair (U,F) should satisfy

$$\nabla F(w) = \nabla U(w) \nabla f(w). \tag{2.10}$$

We prove that (2.5) is a necessary and sufficient condition for the existence of a large family of entropies for (2.1).

Lemma 2.1. *Consider the compressible Euler system* (2.1). *Let us view C as a function of the variables* (p, S).

(1) *Suppose first that $\partial_S C$ is identically zero, so* (2.5) *holds. Then the mathematical entropies U are of the form*

$$U(u, p, S) = \rho h(S) + \rho h_1(u, p) \tag{2.11a}$$

(*with ρ viewed as a function of (p, S)*), *where h is an arbitrary function and h_1 satisfies the following wave equation:*

$$\partial^2_{uu} h_1 - \partial_p \big(C_0(p)^2 \, \partial_p h_1 \big) = 0. \tag{2.11b}$$

(2) *Suppose that $\partial_S c$ does not vanish identically on any interval. Then the mathematical entropies U are of the form*

$$U(u, p, S) = \rho h(S), \tag{2.12}$$

where h is an arbitrary function.

The existence of entropies that depend upon u for particular equations of state was discovered by Schochet [25] and Croisille and Villedieu [11]. For simplicity, we did not include the trivial entropies ρ, ρu, and ρE in case (2) of Lemma 2.1. However, the latter are recovered in case (1) by a suitable choice of h_1.

Lemma 2.2. *Assume that* (2.5) *holds. We denote by g an antiderivative of $-1/C_0(p)^2$ and by G an antiderivative of pg', i.e.,*

$$g'(p) = -\frac{1}{C_0(p)^2}, \qquad G'(p) = p g'(p).$$

Then there exist two functions $g_1 = g_1(S)$ and $g_2 = g_2(S)$, one of which is nonconstant, that satisfy

$$g_1'(S) \geq 0 \quad \text{and} \quad g_2'(S) \geq 0 \tag{2.13}$$

so that

$$\frac{1}{\rho(p, S)} = g(p) + g_1(S), \qquad e(p, S) = -G(p) + g_2(S). \tag{2.14}$$

The assumption (2.6) made in Theorem 2.1 corresponds to the choice $g_1(S) = aS$ and $g_2(S) = bS$ in (2.14). We continue our discussion with the more general assumption, i.e., (2.5). We shall return to (2.6) in the course of the proof of Theorem 2.1. Assuming (2.5), we now show that (2.1) can be formally decoupled into a system of two equations for u and p and a scalar equation for S. As explained now, the decoupling holds for smooth solutions in Lagrangian coordinates but also in a weaker sense for weak solutions in Eulerian coordinates.

For smooth solutions, system (2.1) is equivalent to

$$\begin{aligned} \partial_t p + u \partial_x p + \frac{C_0(p)^2}{\rho} \partial_x u &= 0, \\ \partial_t u + u \partial_x u + \frac{1}{\rho} \partial_x p &= 0, \\ \partial_t S + u \partial_x S &= 0. \end{aligned} \tag{2.15}$$

The decoupling becomes clear when using Lagrangian coordinates. Consider the change of variables $(t, x) \to (t, y)$ defined by

$$\partial_t y = \rho u, \qquad \partial_x y = -\rho. \tag{2.16}$$

Again, for smooth solutions, (2.15) transforms into

$$\begin{aligned} \partial_t p + C_0(p)^2 \partial_y u &= 0, \\ \partial_t u + \partial_y p &= 0, \\ \partial_t S &= 0. \end{aligned} \tag{2.17}$$

The first two equations in (2.17) can be solved independently of the last one. This is not true when (2.5) is not satisfied and C_0 depends on S as well.

Next, consider the case of discontinuous solutions. The change of variables $(t, x) \to (t, y)$ is Lipschitz continuous if ρ and u are measurable and bounded functions. We recall that the Lagrangian–Eulerian transformation preserves the notion of entropy weak solution [32]. In view of (2.14), the equations in (2.1) take the form

$$\begin{aligned} \partial_t \frac{1}{g + g_1} + \partial_x \frac{u}{g + g_1} &= 0, \\ \partial_t \frac{u}{g + g_1} + \partial_x \left(\frac{u^2}{g + g_1} + p \right) &= 0, \\ \partial_t \frac{-G(p) + g_2(S) + u^2/2}{g + g_1} + \partial_x \left(\frac{(-G(p) + g_2(S) + u^2/2)u}{g + g_1} + pu \right) &= 0. \end{aligned} \tag{2.18}$$

Passing to Lagrangian coordinates, we are left with

$$\begin{aligned} \partial_t (g(p) + g_1(S)) - \partial_y u &= 0, \\ \partial_t u + \partial_y p &= 0, \\ \partial_t \left(-G(p) + \frac{u^2}{2} + g_2(S) \right) + \partial_y (pu) &= 0, \end{aligned} \tag{2.19}$$

i.e.,

$$\begin{aligned} \partial_t g(p) - \partial_y u &= -\partial_t g_1(S), \\ \partial_t u + \partial_y p &= 0, \\ \partial_t \left(-G(p) + \frac{u^2}{2} \right) + \partial_y (pu) &= -\partial_t g_2(S). \end{aligned} \tag{2.20}$$

Observe that the right-hand side of (2.20) depends on u and p but not on S. Consider the first two equations in (2.17) and (2.20). The left-hand sides in (2.17) and (2.20) coincide (up to multiplying the first equation in (2.17) by $g'(p)$); the right-hand side of (2.17) is identically zero, whereas the right-hand side of (2.20) contains a term due to the fact that the specific entropy need not be a smooth function.

Possible oscillations in the sequence $(u_\varepsilon, p_\varepsilon, S_\varepsilon)$ are classically described by a Young measure $\nu_{t,x}$, i.e., a probability measure for almost every (t, x) such that

$$f(u_\varepsilon, p_\varepsilon, S_\varepsilon) \rightharpoonup \langle \nu_{t,x}, f \rangle \quad L^\infty \text{ weak-}\star$$

for every continuous function f. Uniform bounds on the amplitude of the $(u_\varepsilon, p_\varepsilon, S_\varepsilon)$'s are assumed for the time being (cf. Lemma 2.4 below). The following lemma is concerned with the reduction of the Young measure to a Dirac mass measure.

Lemma 2.3. *Let $\nu = \nu(u, p, S)$ be a Young measure with compact support that satisfies Tartar's commutation relations*

$$\langle \nu, U_1 F_2 - U_2 F_1 \rangle = \langle \nu, U_1 \rangle \langle \nu, F_2 \rangle - \langle \nu, U_2 \rangle \langle \nu, F_1 \rangle \tag{2.21}$$

for any (U_1, F_1) and (U_2, F_2) in the following family of entropy pairs:

$$\begin{aligned} U(u, p, S) &= \rho h(S) + \rho h_1(u, p), \\ F(u, p, S) &= \rho u h(S) + \rho u h_1(u, p) + k_1(u, p), \end{aligned} \tag{2.22}$$

where h is arbitrary, h_1 is a solution to (2.11b), *and k_1 is given by solving $k_{1,u} = C_0(p)^2 h_{1,p}$ and $k_{1,p} = h_{1,p}$. Suppose that either the closure of the support of ν contains no point with $\rho = 0$ or C_0 is defined from the equation of state* (2.8b). *Then ν is a tensor product of the form*

$$\nu(u, p, S) = \delta_{u_*} \otimes \delta_{p_*} \otimes \mu(S), \tag{2.23}$$

where u_* and p_* are constants and $\mu = \mu(S)$ is a probability measure.

Let us complete Lemma 2.3 with additional comments. We recall that Tartar obtained a reduction of the Young measure in the case of scalar equations using only one entropy function. Next, DiPerna proved a similar result for the elasticity system [13]. Following [13] and [8, 31], it is not difficult to check the following result for the 3×3 system. If the Young measure ν has a small support in the variables (u, p) about a constant state (u_*, p_*), then one extra entropy pair is enough to obtain the reduction of the Young measure to the form (2.23). Namely, one can consider the mathematical entropy

$$U(u, p, S) = \frac{u g(p)}{g(p) + g_1(S)}, \qquad F(u, p, S) = p + \frac{u^2}{2} \frac{g(p) - g_1(S)}{g(p) + g_1(S)}. \tag{2.24}$$

It remains to derive the required a priori estimates. The following result was first mentioned in [27].

Lemma 2.4. *Each domain*

$$\left\{ \pm u + \int^p \frac{dp}{C_0(p)} \leq \text{const.} \right\}$$

is an invariant domain for (2.1) *and* (2.3). *The uniform bounds*

$$\begin{aligned} &0 \leq p^\varepsilon(t, x), |u^\varepsilon(t, x)| \leq \text{const.}, \\ &0 \leq \sup_{x \in \mathbb{R}} S^\varepsilon(t, x) \leq \sup_{x \in \mathbb{R}} S^\varepsilon(0, x) \end{aligned} \tag{2.25}$$

hold for the approximate solutions $(u^\varepsilon, p^\varepsilon, S^\varepsilon)$ generated by (2.3) *with a constant independent of t, x, and ε.*

We now give the proofs of the lemmas stated above.

Proof of Lemma 2.1. It is convenient to use the Lagrangian coordinates (2.16) which transform (2.1) into

$$\begin{aligned} \partial_t u + \partial_y p &= 0, \\ \partial_t p + C(p, S)^2 \partial_y u &= 0, \\ \partial_t S &= 0. \end{aligned}$$

Searching for conservation laws of the form $\partial_t\varphi(u,p,S)+\partial_y\psi(u,p,S)=0$, we arrive at the following necessary and sufficient conditions:

$$\begin{aligned}\psi_u &= C(p,S)^2\varphi_p,\\ \psi_p &= \varphi_u,\\ \psi_S &= 0.\end{aligned}\tag{2.26}$$

Eliminating ψ in (2.26), we obtain

$$\begin{aligned}\varphi_{uu} &= \big(C(p,S)^2\varphi_p\big)_p,\\ \varphi_{uS} &= 0,\\ (C(p,S)^2\varphi_p)_S &= 0.\end{aligned}\tag{2.27}$$

First, suppose that $C=C_0(p)$. Then the third equation in (2.27) reduces to $\varphi_{pS}=0$. It follows that φ has the form

$$\varphi(u,p,S)=h(S)+h_1(u,p),$$

where h is an arbitrary function of S and h_1 is a solution to the wave equation (2.11b). The associated entropy flux ψ is obtained by integration from (2.26). We can recover the conservation laws (2.1) with suitable choices of h and h_1. We also observe that the entropy flux does not depend on S. In particular, this implies that the pair (h_1,ψ) is also an entropy pair for the 2×2 system composed of the first two equations in (2.17).

Next, consider the case that $C_S\neq 0$. Taking the u-derivative of the third equation in (2.27) and combining it with the second equation in (2.27) yields $(C^2)_S\varphi_{up}=0$. Therefore, φ can be decomposed into

$$\varphi(u,p,S)=h_2(p,S)+h_3(u).$$

From the first equation in (2.27), it follows that $h_3''(u)=\partial_p\big(C^2\partial_p h_2\big)(p,S)$, which therefore equals some constant, say α. We deduce the general form of h_2 and h_3 by integrating out the latter:

$$\varphi(u,p,S)=\frac{\alpha u^2}{2}+\beta u+\int^p\frac{\alpha p+\gamma}{C^2(p,S)}dp+h(S),$$

where α, β, and γ are arbitrary constants and h is an arbitrary function of S. Choosing the constant in a proper way, we see that the only available entropies are the specific volume, the velocity, the energy, and any function of the physical entropy S. In other words, there is no extra conservation law but

$$\partial_t S=0.$$

The above conclusions are easily rewritten in the Eulerian setting by observing that $U=\rho\varphi$, $F=\rho u\varphi+\psi$ is an entropy pair for (2.1) if (φ,ψ) is an entropy pair in the Lagrangian framework. Finally, the conditions that guarantee the convexity of the entropies are derived from straightforward computations; cf. also Harten [16] for the entropy (2.12). The proof of Lemma 2.1 is complete. ∎

Proof of Lemma 2.2. Using the specific volume $v=1/\rho$ instead of ρ, condition (2.5) reads $-p_v=C_0(p)^2$, which is readily integrated (in view of the definition of g, i.e., $g'(p) = -1/C_0(p)^2$) into $g(p) = v - g_1(S)$ for some function g_1. This leads to the expression for the density ρ in (2.14). The expression for the energy e follows immediately by integration from the constraint $\tilde{e}_p + p\tilde{v}_p = 0$. Observe that all of standard thermodynamics constraints are satisfied. In particular, we have

$$p\,dv + de = (pg_1'(S) + g_2'(S))\,dS \equiv TdS$$

with $T > 0$ if (2.13) holds and if not both of the functions g_1 and g_2 are constants. □

Proof of Lemma 2.3. First, consider relation (2.21) for the family of entropies of the form of (2.11) with $h = 0$. We observe that the latter reduces to the classical commutation relations for the isentropic system. Namely, define a two-variable Young measure $\tilde{\nu} = \tilde{\nu}(u,p)$ by the relation

$$\langle \tilde{\nu}, \theta \rangle = \frac{\langle \nu, \rho(p,S)\theta(u,p)\rangle}{\langle \nu, \rho(p,S)\rangle} \qquad \text{for every continuous } \theta.$$

For every continuous h_1 and h_2, we have

$$\begin{aligned} &\langle \nu, \rho h_1(\rho u h_2 + k_2) - \rho h_2(\rho u h_1 + k_1)\rangle \\ &\quad = \langle \nu, \rho h_1\rangle\langle \nu, (\rho u h_2 + k_2)\rangle - \langle \nu, \rho h_2\rangle\langle \nu, (\rho u h_1 + k_1)\rangle. \end{aligned} \tag{2.28}$$

Using (2.28) and choosing $h_2 = 1$ and $k_2 = 0$, we get the following relation for the new measure $\tilde{\nu}$:

$$\langle \tilde{\nu}, \rho u h_1\rangle - \langle \tilde{\nu}, (\rho u h_1 + k_1)\rangle = \langle \tilde{\nu}, h_1\rangle\langle \tilde{\nu}, u\rangle\langle \nu, \rho\rangle - \langle \nu, \rho\rangle \left\langle \tilde{\nu}, uh_1 + \frac{k_1}{\rho} \right\rangle.$$

This combined with (2.28) easily yields

$$\langle \tilde{\nu}, h_1 k_2 - k_2 h_1\rangle = \langle \tilde{\nu}, h_1\rangle\langle \tilde{\nu}, k_2\rangle - \langle \tilde{\nu}, k_2\rangle\langle \tilde{\nu}, h_1\rangle. \tag{2.29}$$

Therefore, we can apply the classical reduction theorem to $\tilde{\nu}$ and (2.29), that is, either the result in [13] in the strictly hyperbolic case (if the support of ν does not meet the vaccum) or the result in [14] if the equation of state is given by (2.8b) and a vacuum is allowed. It follows that the Young measure $\tilde{\nu}$ is either a Dirac mass in (u,p) or has its support on the vacuum line $\{p = 0\}$. We observe that we may modify the measure ν at the vacuum points to meet our purpose, and (2.23) is established. □

Proof of Lemma 2.4. Under assumption (2.5), system (2.1) admits the three Riemann invariants $\pm u + \int^p dp/C_0(p)$ and S as is clear from the decoupled formulation (2.17). The desired estimates in Lemma 2.4 are a consequence of Chuey, Conley, and Smoller's theory about bounded invariant domains for nondegenerate parabolic equations [9]. The relevant properties for each of the above functions, say $G(u,p,S) = G(\rho, \rho u, \rho E)$, are as follows:

$$\nabla_{(\rho,\rho u,\rho E)} G \text{ is a left eigenvector} \tag{2.30a}$$

of the Jacobian matrix of the conservative system (2.1), and G is a quasi-convex function, that is, for every $\xi \in \mathbb{R}^3$,

$$\xi \cdot \nabla_{(\rho,\rho u,\rho E)} G = 0 \Longrightarrow \nabla^2_{(\rho,\rho u,\rho E)} G \geq 0. \tag{2.30b}$$

These properties are easily checked for $G = \pm u + \int^p dp/C_0(p)$ or $G = S$.

Note that assumption (2.5) is not necessary to derive the a priori estimate for S^ε; cf. [3] for details. Consider for simplicity a sequence of entropy solutions to (2.1). Consider the entropy inequality associated with the entropy (2.12) for functions satisfying $h' > 0$ and $h'' > 0$. Integrating the inequality $\partial_t \rho^\varepsilon\, h(S^\varepsilon) \le 0$ with respect to the space variable, we get

$$\frac{d}{dt}\int_{\mathbb{R}} \rho^\varepsilon\, h(S^\varepsilon)\, dx \,\le 0. \tag{2.31}$$

Writing (2.31) for the family of functions $h_q(S) = S^q$ with $q \to \infty$ yields at the limit

$$\frac{d}{dt}\sup_{x\in\mathbb{R}} S^\varepsilon(t,x) \le 0,$$

at least away from the vacuum. We observe that without loss of generality the value of S can be arbitrarily modified at a point where $\rho = 0$. Namely, in the equations in (2.1) and the entropy inequalities, all of the terms containing the variable S are multiplied by ρ. The proof of Lemma 2.4 is complete. □

Proof of Theorem 2.1. In view of Lemma 2.4, there exists a Young measure to represent the composite weak-$\star$ limits of the sequence $(u^\varepsilon, p^\varepsilon, S^\varepsilon)$. In order to apply the compensated compactness method to (2.3), we recall [13] that the entropy dissipation measures corresponding to the entropies in (2.22),

$$\partial_t U(u,p,S) + \partial_x F(u,p,S),$$

remain in a compact subset of the negative Sobolev space H^{-1}_{loc}. This follows from an energy-type estimate for (2.3) and Murat's lemma [21]. Therefore, the classical div–rot lemma applies and provide us with (2.21). Lemma 2.3 shows that the Young measure is a Dirac mass in the variables (u,p) but not necessarily in S. It follows that there is a pair (u,p) and a subsequence such that $(u^\varepsilon, p^\varepsilon)$ converges strongly in all L^q $(q < \infty)$ and S^ε converges weakly. Extracting another subsequence if necessary, we can assume that ρ^ε converges weakly to some function, say ρ. We then *define* S by the relation $\tilde\rho(p,S) = \rho$. We emphasize that with the choice of variables (u,p,ρ), the conservative variables and fluxes in (2.1) are linear functions of ρ. Namely, this is obvious for the conservation laws of mass and momentum. On the other hand, for the energy equation, we observe that assumption (2.6) used with (2.14) implies

$$\rho E = \rho\left(\frac{u^2}{2} - G(p) - \frac{bg(p)}{a}\right)$$

so that the conservative variable and the flux in the conservation law of energy are linear in ρ. Note that assumption (2.6) is required in this step of the proof only. Since u^ε and p^ε converge strongly and ρ^ε converges weakly but we need to deal with *linear functions* of ρ^ε only, one can justify the passage to the limit in (2.3) and get (2.1). The proof of Theorem 2.1 is complete. □

3. Relaxed model for two-phase mixtures. Let us begin with the following system of four conservation laws $(i = 1, 2)$:

$$\begin{aligned} &\partial_t \rho_i + \partial_x(\rho_i u_i) = 0,\\ &\partial_t(\rho_i u_i) + \partial_x(\rho_i u_i^2 + p_i) = s_i\,\frac{\rho_2}{\delta}(u_1 - u_2), \end{aligned} \tag{3.1}$$

which describes a mixture of two fluids, say of gas and water bubbles. Here $s_1 = -1$ and $s_2 = 1$. The main unknowns are the mass density ρ_i and the velocity u_i of the phases $i = 1, 2$. We supplement the system with the equations of state for the pressures p_i. We shall assume that

$$p_i = \tilde{p}_i(\rho_i), \quad \frac{d\tilde{p}_i}{d\rho} > 0, \quad \frac{d^2\tilde{p}_i}{d\rho^2} > 0,$$

which ensures the hyperbolicity and genuine nonlinearity of (3.1). The internal energy $e_i = \tilde{e}_i(\rho_i)$ is defined by the relation

$$\frac{d\tilde{e}_i'}{d\rho} = \frac{\tilde{p}_i}{\rho_i^2} \tag{3.2}$$

and the total energy for phase i is $E_i = e_i + u^2/2$. In view of

$$\partial_t(\rho_i E_i) + \partial_x(\rho_i u_i E_i + p_i u_i) = s_i u_i \frac{\rho_2}{\delta}(u_1 - u_2), \tag{3.3}$$

which follows from (3.1) (at least for classical solutions), it is natural to supplement (3.1) with the entropy inequality

$$\partial_t(\rho_1 E_1 + \rho_2 E_2) + \partial_x(\rho_1 u_1 E_1 + \rho_2 u_2 E_2 + p_1 u_1 + p_2 u_2) = -\frac{\rho_2}{\delta}(u_1 - u_2)^2 \le 0. \tag{3.4}$$

This model describes a mixture of gas and liquid (or solid) dropplets. The source term in the momentum equation is called the drag force. The parameter δ can be identified with a relaxation time, i.e., the time for the small dropplets to acquire the same velocity as the one of the gas; cf. Sainsaulieu [24] for the derivation of this model from a more microscopic physical description.

Our focus here is on analyzing the zero-relaxation limiting system derived from (3.1). When $\delta \to 0$, we formally have

$$u_1 = u_2 \equiv u \quad \text{(average flow velocity)}, \tag{3.5}$$

and we are left with the 3 × 3 relaxed system

$$\begin{aligned} \partial_t \rho_i + \partial_x(\rho_i u) = 0, \quad i = 1, 2 \\ \partial_t\big((\rho_1 + \rho_2)u\big) + \partial_x\big((\rho_1 + \rho_2)u^2 + p_1 + p_2\big) = 0, \end{aligned} \tag{3.6}$$

while the entropy inequality (3.4) becomes

$$\partial_t\left(\rho_1 e_1 + \rho_2 e_2 + \frac{(\rho_1 + \rho_2)u^2}{2}\right) + \partial_x\big((\rho_1 e_1 + \rho_2 e_2)u + (p_1 + p_2)u\big) \le 0. \tag{3.7}$$

The main unknowns of the new system are taken to be ρ_1, ρ_2, and u.

At this stage it is interesting to observe that—for classical solutions—(3.6)–(3.7) coincides with the gas dynamics system studied in section 2. However, the weak solutions of both systems do not coincide. To make our point, we introduce the new variables

$$\rho = \rho_1 + \rho_2, \qquad S = \frac{\rho_2}{\rho}$$

together with the familiar notation

$$p = p_1 + p_2, \qquad e = e_1 + e_2, \qquad E = e + \frac{u^2}{2}$$

so that (3.6)–(3.7) takes the form

$$\begin{aligned} \partial_t \rho + \partial_x(\rho u) &= 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + p) &= 0, \\ \partial_t(\rho S) + \partial_x(\rho S u) &= 0 \end{aligned} \tag{3.8}$$

and

$$\partial_t(\rho E) + \partial_x(\rho u E + pu) \leq 0. \tag{3.9}$$

For smooth solutions, (3.8) is equivalent to system (2.1). In particular, the family of mathematical entropies derived in Lemma 2.1 for (2.1) is the same for (3.8). The entropy weak solutions to both systems are distinct, however, since the classical formulation of the gas dynamic system conserves the total energy and lets the entropy decrease, while (3.8)–(3.9) conserves the entropy while decreasing the energy!

We shall prove that the new formulation (3.8)–(3.9) allows us to apply the compensated compactness method in all generality without assuming restriction (2.5) on the equation of state. For the sake of generality, we disregard special form of p that could be derived from the above relation $p = p_1 + p_2$. Consider the approximate solutions ρ^ε, u^ε, and S^ε given by solving

$$\begin{aligned} \partial_t \rho + \partial_x(\rho u) &= \varepsilon\, \partial_{xx} \rho, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + p) &= \varepsilon\, \partial_{xx}(\rho u), \\ \partial_t(\rho S) + \partial_x(\rho S u) &= \varepsilon\, \partial_{xx}(\rho S) \end{aligned} \tag{3.10}$$

from the initial data ρ_0, u_0, and S_0 in L^∞. Observe that inequality (3.9) is automatically satisfied in the limit. For simplicity, we restrict ourselves to solutions that are bounded away from the vacuum. We denote by $\mathrm{BV}(\mathbb{R})$ the space of all Lebesgue-measurable, scalar-valued functions of bounded variation in one variable.

THEOREM 3.1. *Suppose that ρ^ε, u^ε, and S^ε are smooth solutions to* (3.10) *that satisfy the uniform bound*

$$\begin{gathered} 0 < \text{const.} \leq \rho^\varepsilon(t,x) \leq \text{const.}, \\ |u^\varepsilon(t,x)| \leq \text{const.}, \qquad |S^\varepsilon(t,x)| \leq \text{const.}, \end{gathered} \tag{3.11}$$

where the constants are independent of t, x, and ε. Suppose that the initial data S_0 belongs to the space $\mathrm{BV}(\mathbb{R})$. *Then there is a triplet (ρ, u, S) in $L^\infty(\mathbb{R}_+ \times \mathbb{R})^3$ and a subsequence of $(\rho^\varepsilon, u^\varepsilon, S^\varepsilon)$ so that*

$$\rho^\varepsilon \to \rho, \qquad u^\varepsilon \to u, \qquad S^\varepsilon \to S \tag{3.12}$$

in the strong L^q norm for all finite q, and (ρ, u, S) is an entropy weak solution to (3.8)–(3.9).

This convergence and existence result can be extended to the cases where either one of the conservation laws of mass, momentum, and energy is used as an "entropy inequality," while the rest of the equations are written as conservation laws (including the specific entropy). In other words, Theorem 3.1 holds for any of the following two formulations:

$$\begin{aligned} \partial_t\rho + \partial_x(\rho u) &= 0, & \partial_t\rho + \partial_x(\rho u) &\le 0,\\ \partial_t(\rho u) + \partial_x(\rho u^2 + p) &\le 0, & \partial_t(\rho u) + \partial_x(\rho u^2 + p) &= 0,\\ \partial_t(\rho E) + \partial_x(\rho u E + pu) &= 0, & \partial_t(\rho E) + \partial_x(\rho u E + pu) &= 0,\\ \partial_t(\rho S) + \partial_x(\rho S u) &= 0, & \partial_t(\rho S) + \partial_x(\rho S u) &= 0. \end{aligned} \tag{3.13}$$

Proof of Theorem 3.1. The heart of the proof is the derivation of an a priori estimate for the total variation of S^ε. A straightforward calculation using the first and third equations in (3.11) shows that S is a solution to the scalar advection-diffusion equation

$$\partial_t S^\varepsilon + V^\varepsilon\,\partial_x S^\varepsilon = \varepsilon\partial_{xx}S^\varepsilon \tag{3.14}$$

with

$$V^\varepsilon = u^\varepsilon - 2\varepsilon\partial_x(\log\rho^\varepsilon).$$

We use a classical approach to estimate the total variation of S as follows. Set $w = \partial_x S^\varepsilon$ and take $q > 1/2$. From (3.14), we deduce

$$\partial_t|w|^{2q} + \partial_x(V^\varepsilon|w|^{2q}) + (2q-1)|w|^{2q}\partial_x V^\varepsilon \le \varepsilon\partial_{xx}\left(\frac{|w|^{2q+1}}{(2q+1)}\right).$$

Letting $q \to 1/2$, we obtain

$$\partial_t|w| + \partial_x(V^\varepsilon|w|) \le \varepsilon\partial_{xx}\left(\frac{|w|^2}{2}\right).$$

Integrating the latter over $x \in \mathbb{R}$ yields an estimate for $\partial_x S^\varepsilon$:

$$\int_{\mathbb{R}} |\partial_x S^\varepsilon(t,x)|\,dx \;\le\; \int_{\mathbb{R}} |S_0'(x)|\,dx \tag{3.15}$$

when S_0 belongs to $W^{1,1}(\mathbb{R})$. The case where $S_0 \in \mathrm{BV}(\mathbb{R})$ is treated by a standard regularization process. By Helly's theorem, the sequence $\{S^\varepsilon\}$ (or a subsequence of it) converges in any L^q norm, $1 \le q < \infty$.

Next, we will apply the compensated compactness method to (3.10). We shall use functions that are entropy functions for the isentropic gas dynamics system but not for the full system; here we follow the method proposed by Benzoni-Gavage and Serre. We present the main lines of the proof and refer to [1] for details. The total variation estimate above allows us to control the nonconservative terms that arise. As before, let C denote the Lagrangian sound speed and consider the class $\mathcal{A}$ of all pairs (U,F) of the form $U = \rho h_1(u,p)$ that satisfies the wave equation (2.11b) with C_0 replaced by C, i.e.,

$$\partial^2_{uu}h_1 - \partial_p\big(C(p,S)^2\,\partial_p h_1\big) = 0 \tag{3.16}$$

for each value of S. For an entropy pair (U,F) in the class $\mathcal{A}$, it is not hard to check using (3.15) that

$$\partial_t U + \partial_x F \in \text{compact set of } H^{-1}_{\mathrm{loc}}(\mathbb{R}_+\times\mathbb{R})$$

so that the commutation equation (2.21) is satisfied by every entropy pairs in $\mathcal{A}$. From the reduction theorem of DiPerna [13] for genuinely nonlinear systems, we deduce that u^ε and p^ε converge in a strong sense. The proof of Theorem 3.1 is complete. □

Remark 3.1. Chuey, Conley, and Smoller's theory does not apply to system (3.10), and there are in general no available invariant regions. In the special case where $p = a^2 S\rho/(1-\rho)$, Peng [23] shows the existence of invariant regions for the Riemann problem and obtains global existence using the Glimm scheme.

It would be interesting to show that as the relaxation parameter δ tends to 0, the solutions of (3.1) converge strongly to a solution of the relaxed system (3.8). One can anticipate that the relaxation terms have the same effect in canceling oscillations, as we proved for the vanishing viscosity method. Rigorous convergence results were established by Chen, Levermore, and Liu [6, 7] for genuinely nonlinear 2×2 systems with relaxation. Their argument is based on the fact that from every convex entropy of the relaxed equations stems a *compatible* entropy of the full system. The scalar product of the entropy gradient with the relaxation term has a constant sign, which yields the required $H^{-1}_{\rm loc}$ compactness property.

Even though (3.1) consists of two genuinely nonlinear systems, coupled only through the lower-order source terms, a rigorous proof of convergence of the solutions $(\rho_i^\delta, u_i^\delta)_{i=1,2}$ to a solution of the relaxed system (3.8)–(3.9) seems extremely difficult. This is due mainly to the form of the relaxation terms, as we now explain.

Let us begin with the vanishing viscosity regularization for (3.1), i.e., for $i = 1, 2$,

$$\begin{aligned} &\partial_t \rho_i + \partial_x(\rho_i u_i) = \varepsilon\, \partial_{xx}\rho_i, \\ &\partial_t(\rho_i u_i) + \partial_x(\rho_i u_i^2 + p_i) = s_i\, \frac{\rho_2}{\delta}(u_1 - u_2) + \varepsilon\, \partial_{xx}(\rho_i u_i). \end{aligned} \tag{3.17}$$

We show in Proposition 3.1 below that (3.17) does not admit any invariant domain in the sense of Chuey, Conley, and Smoller. Nevertheless, assuming a uniform L^∞ bound on the solutions of (3.17), a standard energy method yields the following a priori estimate:

$$\delta^{-1/2}\|u_1^{\varepsilon,\delta} - u_2^{\varepsilon,\delta}\|_{L^2(\mathbb{R})} + \varepsilon^{1/2}\Sigma_{i=1,2}\big(\|\partial_x \rho_i^{\varepsilon,\delta}\|_{L^2(\mathbb{R})} + \|\partial_x u_i^{\varepsilon,\delta}\|_{L^2(\mathbb{R})}\big) \le \text{const.} \tag{3.18}$$

Letting δ tend to 0 first, it is easy to see that the solutions of (3.17) converge strongly in $L^2_{\rm loc}$ to a solution of (3.8)–(3.9), and Theorem 3.1 applies to the resulting 3×3 system.

On the other hand, if we let ε tend to 0 first, the desired convergence result, and therefore the existence of entropy weak solutions to (3.1) and (3.4), is a consequence of DiPerna's theorem for 2×2 genuinely nonlinear systems [14]. Furthermore, for each $\delta > 0$, these solutions satisfy the entropy inequality (3.4). On the other hand, in the terminology of [6, 7], the total energy is a compatible entropy for (3.1). It turns out that there is no other compatible entropy.

PROPOSITION 3.1.

(1) *System* (3.17) *does not admit any invariant domain in the sense of Chuey, Conley, and Smoller.*

(2) *Every entropy of* (3.17) *takes the form* $\eta_1(\rho_1, u_1) + \eta_2(\rho_2, u_2)$, *where* η_i *is an entropy for the associated subsystem*

$$\begin{aligned} &\partial_t \rho_i + \partial_x(\rho_i u_i) = 0, \\ &\partial_t(\rho_i u_i) + \partial_x(\rho_i u_i^2 + p_i) = 0. \end{aligned} \tag{3.19}$$

(3) *The only compatible entropy for* (3.17) *is the total energy* $\rho_1 E_1 + \rho_2 E_2$.

Proof. Let R_δ denote the right-hand side of (3.1). To show the first claim, observe that the invariant domains in the sense of [9] must be in the form $\pm w_i(\rho_1, \rho_1 u_1, \rho_2, \rho_2 u_2) \leq 0$, where $(w_i)_{i=1,4}$ are four Riemann invariants for (3.1). Another necessary condition is that

$$\nabla w_i \cdot R_\delta \leq 0. \tag{3.20}$$

An easy calculation shows that the Riemann invariants of (3.17) are the same as those of the subsystems and thus depend only on either (ρ_1, u_1) or (ρ_2, u_2), whereas R_δ is proportional to $u_1 - u_2$. Thus (3.20) cannot hold.

We now examine the structure of the entropy pairs $(U, F)(\rho_1, u_1, \rho_2, u_2)$ for (3.17). Eliminating the entropy flux F in condition (2.10) yields the following equations for U:

$$\begin{bmatrix} (u_1 - u_2) & -\frac{\tilde{p}_2'}{\rho_2} & \frac{\tilde{p}_1'}{\rho_1} & 0 \\ -\rho_2 & (u_1 - u_2) & 0 & -\frac{\tilde{p}_2'}{\rho_2} \\ \rho_1 & 0 & (u_1 - u_2) & \frac{\tilde{p}_1'}{\rho_1} \\ 0 & \rho_1 & -\rho_2 & (u_1 - u_2) \end{bmatrix} \begin{bmatrix} U_{\rho_1,\rho_2} \\ U_{\rho_1,u_2} \\ U_{u_1,\rho_2} \\ U_{u_1,u_2} \end{bmatrix} = 0 \tag{3.21}$$

together with

$$U_{u_i,u_i} = \frac{\tilde{p}_i'}{\rho_i} U_{\rho_i,\rho_i}, \quad i = 1, 2. \tag{3.22}$$

If $u_1 \pm (p_1')^{1/2} \neq u_2 \pm (p_2')^{1/2}$, the matrix in (3.21) is invertible and therefore

$$U_{\rho_1,\rho_2} = U_{\rho_1,u_2} = U_{u_1,\rho_2} = U_{u_1,u_2} = 0.$$

The conditions in (2.10) induce exactly the same constraints on F. Thus the entropy pairs for (3.1) have the form $(U, F) = \big(\eta_1(\rho_1, u_1) + \eta_2(\rho_2, u_2), q_1(\rho_1, u_1) + q_2(\rho_2, u_2)\big)$, where η_i satisfies (3.22), i.e., is an entropy for the ith subsystem, and q_i is the associated entropy flux. For such a pair $(U, F) = (\eta_1 + \eta_2, q_1 + q_2)$ with convex η_1 and η_2, the solutions of (3.1) obtained by the viscosity method satisfy

$$\partial_t U(\rho_1^\delta, u_1^\delta, \rho_2^\delta, u_2^\delta) + \partial_x F(\rho_1^\delta, u_1^\delta, \rho_2^\delta, u_2^\delta) \leq \nabla U \cdot R_\delta. \tag{3.23}$$

We now seek compatible entropies, i.e., pairs (U, F) for which the term on the right-hand side of (3.23) is nonpositive. In other words, we require that

$$\operatorname{sign}(u_1 - u_2)(g_1 - g_2) \leq 0 \tag{3.24}$$

for all quadruples $(\rho_1, u_1, \rho_2, u_2)$, where $g_i(\rho_i, u_i) = \eta_{i,u}(\rho_i, u_i)/\rho_i$. Condition (3.24) implies that g_i depends only on u_i, and thus

$$\eta_{i,u} = \rho_i g_i(u_i). \tag{3.25}$$

From (3.22) and the definition (3.2) of the internal energy, we obtain

$$\frac{\partial^2 \eta_i}{\partial \rho^2} = \frac{\tilde{p}_i'}{\rho_i^2} \frac{\partial^2 \eta_i}{\partial u^2} = \tilde{e}_i'' g_i''(u_i).$$

Integrating twice with respect to ρ_i, we find

$$\eta_i = \tilde{e}_i g_i'(u_i) + \rho_i h(u_i) + k(u_i)$$

for some functions h and k. From (3.25), we see that k is a constant and

$$\rho_i^{-1}\tilde{e}_i g_i''(u_i) = \rho_i^{-1}\tilde{e}_i(\rho_i) g_i''(u_i) = g_i(u_i) - h(u_i),$$

which implies that $g_i'' = 0$ since the right-hand side is independent of ρ_i. Thus g_i is a linear function of u_i, while (3.25) again implies that $h' = g_i$. We conclude that the only compatible entropies have the form

$$\eta_i = \alpha(\rho_i u_i^2/2 + e_i) + \beta\rho_i u_i + \gamma\rho_i + \zeta$$

for some constants α, β, γ, and ζ, which completes the proof of Proposition 3.1. □

REFERENCES

[1] S. Benzoni-Gavage and D. Serre, *Compacité par compensation pour une classe de systèmes hyperboliques de p lois de conservation* ($p \geq 3$), Rev. Mat. Iberoamericana, 10 (1994), pp. 557–579.

[2] E. Bonnetier and P. G. LeFloch, in preparation.

[3] G.-Q. Chen, *The compensated compactness method and the system of isentropic gas dynamics*, preprint, Mathematical Sciences Research Institute, Berkeley, CA, 1991.

[4] G.-Q. Chen and C. M. Dafermos, *The vanishing viscosity method in one-dimensional thermoelasticity*, preprint, Brown University, Providence, RI, 1994.

[5] G.-Q. Chen and P. G. LeFloch, *Entropy kernel and weak solutions of the isentropic Euler equations*, C. R. Acad. Sci. Paris Ser. I Math., submitted.

[6] G.-Q. Chen, D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation term and entropy*, Comm. Pure Appl. Math., 7 (1994), pp. 787–830.

[7] G.-Q. Chen and T.-P. Liu, *Zero relaxation and dissipation limits for hyperbolic conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 755–781.

[8] G.-Q. Chen and Y.-G. Lu, *Convergence of the approximate solutions to isentropic gas dynamics*, Acta Math. Sci., 10 (1990), pp. 39–45.

[9] K. N. Chuey, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 372–411.

[10] R. Courant and K. O. Friedrichs, Supersonic Flows and Shock Waves, Springer-Verlag, New York, 1948.

[11] J.-P. Croisille and P. Villedieu, *Entropies de Lax pour les équations d'Euler en deséquilibre thermochimique*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 723–727.

[12] C. M. Dafermos, *Hyperbolic systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., NATO Adv. Sci. Inst. Ser. C Math Phys. Sci. 111, D. Reidel, Dordrecht, the Netherlands, 1983, pp. 25–70.

[13] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.

[14] R. J. DiPerna, *Convergence of the viscosity method for isentropic gas dynamics*, Comm. Math. Phys., 91 (1983), pp. 1–30.

[15] R. J. DiPerna, Compensated compactness and general systems of conservation laws, Trans. Amer. Math. Soc., 292 (1985), pp. 383–421.

[16] A. Harten, *On the symmetric form of systems of conservation laws with entropy*, J. Comput. Phys., 49 (1983), pp. 151–164.

[17] P. D. Lax, *Shock wave and entropy*, in Contributions to Functional Analysis, E. A. Zarantonello, ed., Academic Press, New York, 1971, pp. 603–634.

[18] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM, Philadelphia, 1973.

[19] P.-L. Lions, B. Perthame, and P. Sougadinis, *Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates*, preprint, Comm. Pure Appl. Math., to appear.

[20] P.-L. Lions, B. Perthame, and E. Tadmor, *Kinetic formulations for the p-system and Euler system*, Comm. Math. Phys., 163 (1994), pp. 415–431.

[21] F. Murat, *L'injection du cône positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout* $q < 2$, J. Math. Pures Appl., 60 (1981), pp. 309–322.

[22] F. Murat, *A survey on compensated compactness*, in Contributions to Modern Calculus of Variation, L. Cesari, ed., Pitman Research Notes in Math. Series 148, Longman, Harlow, UK, 1987, pp. 145–183.

[23] Y. J. Peng, *Solutions faibles globales pour un modèle d'écoulements diphasiques*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 21 (1994), pp. 523–540.

[24] L. Sainsaulieu, *An Euler system modeling vaporizing sprays*, in Dynamics of Heterogeneous Combustion and Reacting Systems, A. L. Kuhl et al., eds., Progress Astronautics Aeronautics, 152, American Institute of Aeronautics and Astronautics, Washington, DC, 1993, pp. 280–305.

[25] S. Schochet, *Examples of measure-valued solutions*, Comm. Partial Differential Equations, 14 (1989), pp. 545–575.

[26] D. Serre, *La compacité par compensation pour les systèmes de deux equations à une dimension d'espace*, J. Math. Pures Appl., 65 (1986), pp. 423–468.

[27] D. Serre, unpublished notes, cours de diplôme d'études approfondies, University of Paris VI, 1988.

[28] D. Serre, *Richness and the classification of quasilinear hyperbolic systems*, in Multidimensional Hyperbolic Problems and Computations, IMA Vol. Math. Appl. 29, Springer-Verlag, New York, 1991, pp. 315–333.

[29] D. Serre, *Oscillations nonlinéaires des systèmes hyperboliques: Méthodes et résultats qualitatifs*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 351–417.

[30] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Vol. IV, Pitman Research Notes in Math., Pitman, San Francisco, 1979, pp. 136–212.

[31] I. Vecchi, *Entropy compactification in Lagrangian gas dynamics*, Math. Methods Appl. Sci., 14 (1991), pp. 207–216.

[32] D. Wagner, *Equivalence of Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 516–529, May 1997

THE CAUCHY PROBLEM AND THE CONTINUOUS LIMIT FOR THE MULTILAYER MODEL IN GEOPHYSICAL FLUID DYNAMICS[*]

T. COLIN[†]

Abstract. We study a multilayer model in geophysical fluid dynamics that approximately governs the large-scale motions of the atmosphere or the ocean. The model consists of n two-dimensional Euler equations which represent the evolution of n layers of liquid. These equations are written using the potential vorticity. The potential vorticity in each layer is obtained from the velocity potential of the adjacent layers. We show that the Cauchy problem for this model is globally well-posed in time for smooth initial data. In the second part of the paper, we let the number of layers tend to infinity while their thickness tends to zero. We write the system as a suitable finite-element approximation of a continuous model and show the convergence of this approximation to the classical quasi-geostrophic model.

Key words. geophysical fluid dynamics, layers model, quasi-geostrophic

AMS subject classifications. 35Q35, 35B99, 65M60

PII. S0036141095291269

1. Introduction, setting of the problem, and statement of results.

1.1. The model. Consider a fluid which is formed by the superposition of a finite number n of homogeneous layers with uniform density within each layer, the density being different from one layer to another. The multilayered quasi-geostrophic model describes concervation of potential vorticity ζ_l in each layer with the β-plane approximation. The thickness of each layer has the same value D_0, and we denote by ρ_l the density in the lth layer, $l = 1, \ldots, n$. The nondimensional model (see [9, p. 421]) reads

$$(1)\qquad \begin{cases} \left(\dfrac{\partial}{\partial t} + u_l \dfrac{\partial}{\partial x} + v_l \dfrac{\partial}{\partial y}\right) \zeta_l = -a v_l, \quad l = 1 \text{ to } n, \\ u_l = -\dfrac{\partial \psi_l}{\partial y}, \qquad v_l = \dfrac{\partial \psi_l}{\partial x}, \end{cases}$$

$$(2)\qquad \begin{cases} \zeta_1 = \Delta_{x,y}\psi_1 + f_{r1,2}(\psi_2 - \psi_1), \\ \zeta_l = \Delta_{x,y}\psi_l - f_{rl,1}(\psi_l - \psi_{l-1}) + f_{rl,2}(\psi_{l+1} - \psi_l), \quad 2 \le l \le n-1, \\ \zeta_n = \Delta_{x,y}\psi_n - f_{rn,1}(\psi_n - \psi_{n-1}), \end{cases}$$

where

$$f_{rl,1} = \frac{DF}{D_0} \frac{\rho_0}{\rho_l - \rho_{l-1}}$$

[*] Received by the editors September 5, 1995; accepted for publication (in revised form) January 18, 1996. This research was partially supported by CEA grant DAM/CEL-V W003871/370 (CEA, Centre d'Études de Limeil-Valenton, 94195 Villeneuve St-Georges cedex, France).

http://www.siam.org/journals/sima/28-3/29126.html

[†] Mathématiques Appliquées de Bordeaux (CNRS ERS 123), Université Bordeaux 1, 351 Cours de la Libération, 33405 Talence cedex, France (colin@math.u-bordeaux.fr).

and

$$f_{rl,2} = \frac{DF}{D_0}\frac{\rho_0}{\rho_{l+1} - \rho_l}.$$

In the above equations, (u_l, v_l) is the velocity field within the lth layer. All functions depend only on the horizontal variables x and y and on the time t. The term $-av_l$ in (1) corresponds to the action of the Coriolis force in the β-plane approximation [9]. Note that the function ψ_l is in fact the physical pressure inside the lth layer; see [9, p. 421]. ρ_0 is a characteristic (constant) value for the density of the fluid and $F = f_0^2 L^2/gD$, where L (resp. D), is the characteristic horizontal (resp. vertical) scale of the fluid. f_0 is the Coriolis parameter and g is the gravitational acceleration. We introduce the following notation,

$$\beta_l = \frac{DF\rho_0}{D_0(\rho_l - \rho_{l-1})}\varepsilon^2,$$

where ε is a small (nondimensional) parameter, and we assume that

$$\frac{1}{\delta} \geq \beta_l \geq \delta \quad \text{for } l = 2, \dots, n, \tag{3}$$

where δ is a fixed positive number. Relation (3) is taken to be satisfied independently of n. The parameters $f_{rl,1}$ and $f_{rl,2}$ therefore read

$$f_{rl,1} = \frac{1}{\varepsilon^2}\beta_l, \qquad f_{rl,2} = \frac{1}{\varepsilon^2}\beta_{l+1}.$$

With this notation, system (2) takes the form

$$\begin{cases} \zeta_1 = \Delta_{x,y}\psi_1 + \dfrac{1}{\varepsilon^2}\beta_2(\psi_2 - \psi_1), \\[2ex] \zeta_l = \Delta_{x,y}\psi_l - \dfrac{1}{\varepsilon^2}\beta_l(\psi_l - \psi_{l-1}) + \dfrac{1}{\varepsilon^2}\beta_{l+1}(\psi_{l+1} - \psi_l), \quad 2 \leq l \leq n-1, \\[2ex] \zeta_n = \Delta_{x,y}\psi_n - \dfrac{1}{\varepsilon^2}\beta_n(\psi_n - \psi_{n-1}). \end{cases} \tag{4}$$

A derivation of this model and its physical meaning can be found in [9, pp. 416–422]. See also the appendix of [6].

We will use periodic boundary conditions that correspond to $(x, y) \in \mathbf{T}^2$, where $\mathbf{T}^2$ denotes the two-dimensional torus. The aim of this paper is to show that system (1)–(4) is globally well-posed for smooth initial data and to perform the limit $n \to \infty$ in a suitable sense.

The "natural" limit system is the quasi-geostrophic equation

$$\left(\frac{\partial}{\partial t} - \frac{\partial\psi}{\partial y}\frac{\partial}{\partial x} + \frac{\partial\psi}{\partial x}\frac{\partial}{\partial y}\right)\left(\Delta_{x,y}\psi + \frac{\partial}{\partial z}\left(\beta(z)\frac{\partial\psi}{\partial z}\right)\right) = -a\frac{\partial\psi}{\partial x}. \tag{5}$$

In this equation, z is a spatial variable while the function $z \mapsto \beta(z)$ is proportional to

$$\frac{1}{\dfrac{\partial\rho(z)}{\partial z}},$$

i.e., to the inverse of the square of the Brunt–Väisälä frequency (see [9, pp. 354–358]).

This system is of considerable physical interest because of the qualitative results that are readily deduced from it (see [9] and [11]). For its mathematical treatement, see [1], where the asymptotic expansion leading from some set of primitive equations to (5) is justified, and [2], where the viscous case is considered. See also [7] and [4] for analysis of models without vertical stratification.

1.2. Notation and statement of results. In section 2, we prove that for fixed n, system (1)–(4) is globally well-posed. Our result in this direction is the following (see Theorem 2.1):

Let $s > 2$ and $(\zeta_1^0, \dots, \zeta_n^0) \in (H^s(\mathbf{T}^2))^n$ be such that

$$\sum_{l=1}^{n} \int_{\mathbf{T}^2} \zeta_l^0 = 0.$$

Then there exists a unique solution $(\zeta_1, \dots, \zeta_n) \in (\mathcal{C}(\mathbf{R}^+, H^s(\mathbf{T}^2)))^n$ and $((u_1, v_1), \dots, (u_n, v_n)) \in (\mathcal{C}(\mathbf{R}^+, H^{s-1}(\mathbf{T}^2)))^{2n}$ to (1) and (4) such that $(\zeta_1, \dots, \zeta_n)(t = 0) = (\zeta_1^0, \dots, \zeta_n^0)$.

In the third section, we show how (1)–(4) may be viewed as a finite-element approximation in the vertical direction to the continuous system (5) and how the limit $n \to \infty$ can be performed. Indeed, system (4) can be viewed as a finite-element approximation in only one direction of a continuous three-dimensional elliptic equation. This leads us to introduce the functions $\Psi_1^\varepsilon(x, y, z, t)$, $\Psi_3^\varepsilon(x, y, z, t)$, and $Z_2^\varepsilon(x, y, z, t)$ below in terms of the classical basis functions. However, equation (1) is not a finite-element approximation of the corresponding continuous equation. We therefore need to introduce other auxilliary functions ($\Psi_2^\varepsilon(x, y, z, t)$ and $Z_1^\varepsilon(x, y, z, t)$) in order to write system (1)–(4) as a coherent approximation of the quasi-geostrophic model (5).

Specifically, we denote by $\phi_j^\varepsilon(z)$ the functions defined on $[0, 1]$ such that

$$\begin{aligned}
&\phi_1^\varepsilon(z) = 1 - \frac{z}{\varepsilon} \quad \text{if } 0 \le z \le \varepsilon, \quad 0 \text{ elsewhere},\\
&\phi_i^\varepsilon(z) = 1 - \frac{|z - i\varepsilon|}{\varepsilon} \quad \text{if } z \in [(i-1)\varepsilon, (i+1)\varepsilon], \quad 0 \text{ elsewhere}, \quad i = 2, \dots, n-1,\\
&\phi_n^\varepsilon(z) = 1 - \frac{1 - z}{\varepsilon} \quad \text{if } z \in [1 - \varepsilon, 1], \quad 0 \text{ elsewhere},
\end{aligned}$$

where we assume that $\varepsilon = 1/(n-1)$. We assume that there exists a smooth function $\beta(z)$ such that

$$\beta_l^\varepsilon = \frac{1}{\varepsilon} \int_{(l-2)\varepsilon}^{(l-1)\varepsilon} \beta(z) dz \quad \text{for } l = 2, \dots, n$$

and then introduce

$$\Psi_1^\varepsilon(x, y, z, t) = \sum_{l=1}^{n} \psi_l^\varepsilon(x, y, t) \phi_l^\varepsilon(z). \tag{6}$$

We also construct the piecewise-constant function $\Psi_2^\varepsilon(x, y, z, t)$, which is equal to $\psi_l^\varepsilon(x, y, t)$ on $[(l-1)\varepsilon, l\varepsilon[$ for $l = 1, \dots, n-1$, namely

$$\Psi_2^\varepsilon(x, y, z, t) = \sum_{l=1}^{n-1} \psi_l^\varepsilon(x, y, t) \mathbf{1}_{[(l-1)\varepsilon, l\varepsilon[}(z), \tag{7}$$

where $\mathbf{1}_{[a,b]}(z)$ denotes the characteristic function of $[a,b]$.

In order to solve (4), we need to consider the function Ψ_3^ε of the following form:

$$\Psi_3^\varepsilon = \sum_{i=1}^{n} f_i^\varepsilon \phi_i^\varepsilon. \tag{8}$$

Moreover, we impose the condition that Ψ_3^ε satisfies

$$\int_0^1 \Psi_3^\varepsilon(z)\phi_i^\varepsilon(z)dz = \varepsilon\psi_i^\varepsilon \quad \text{for all } i = 1, \dots, n. \tag{9}$$

Thus (9) reads

$$\int_0^1 \Psi_3^\varepsilon \phi_k^\varepsilon(z)dz = \int_0^1 \sum_{i=1}^{n} f_i^\varepsilon \phi_i^\varepsilon(z)\phi_k^\varepsilon(z)dz = \varepsilon\psi_k^\varepsilon.$$

Using the explicit values of ϕ_l^ε, we find that the coefficients (f_i^ε) are given by

$$\begin{pmatrix} \psi_1^\varepsilon \\ \cdots \\ \psi_n^\varepsilon \end{pmatrix} = B_n \begin{pmatrix} f_1^\varepsilon \\ \cdots \\ f_n^\varepsilon \end{pmatrix}, \tag{10}$$

where B_n is the $n \times n$ matrix

$$B_n = \frac{1}{6}\begin{pmatrix} 2 & 1 & 0 & & & & \\ 1 & 4 & 1 & 0 & & 0 & \\ 0 & \cdots & \cdots & \cdots & 0 & & \\ & 0 & 1 & 4 & 1 & 0 & \\ & & 0 & \cdots & \cdots & \cdots & 0 \\ & 0 & & 0 & 1 & 4 & 1 \\ & & & & 0 & 1 & 2 \end{pmatrix}.$$

In the same direction, we define the piecewise-constant function Z_1^ε by

$$Z_1^\varepsilon(x,y,z,t) = \sum_{l=1}^{n-1} \zeta_l^\varepsilon(x,y,t)\mathbf{1}_{[(l-1)\varepsilon, l\varepsilon[}(z). \tag{11}$$

We also introduce $Z_2^\varepsilon(x,y,z,t)$ defined in the same way as Ψ_3^ε:

$$Z_2^\varepsilon(x,y,z,t) = \sum_{i=1}^{n} g_i^\varepsilon \phi_i^\varepsilon, \tag{12}$$

where (g_i^ε) are given by

$$\begin{pmatrix} \zeta_1^\varepsilon \\ \cdots \\ \zeta_n^\varepsilon \end{pmatrix} = B_n \begin{pmatrix} g_1^\varepsilon \\ \cdots \\ g_n^\varepsilon \end{pmatrix}.$$

With this notation, the original system can be written exactly as follows:

$$\frac{\partial}{\partial t} Z_1^\varepsilon + \nabla \cdot ((U_2^\varepsilon, V_2^\varepsilon)(Z_1^\varepsilon + ay)) = 0, \tag{13}$$

and for all $k = 1, \dots, n$,

$$\int_0^1 Z_2^\varepsilon \phi_k^\varepsilon(z)dz = \int_0^1 \Delta_{x,y}\Psi_3^\varepsilon(z)\phi_k^\varepsilon(z)dz - \int_0^1 \beta(z)\frac{\partial \Psi_1^\varepsilon}{\partial z}\frac{\partial \phi_k^\varepsilon}{\partial z}dz, \tag{14}$$

where

$$U_2^\varepsilon = -\frac{\partial \Psi_2^\varepsilon}{\partial y} \quad \text{and} \quad V_2^\varepsilon = \frac{\partial \Psi_2^\varepsilon}{\partial x}. \tag{15}$$

The result that we obtain is as follows (for a precise statement, see Theorem 3.1):

The sequences Ψ_1^ε, Ψ_2^ε, *and* Ψ_3^ε *converge to the same limit* Ψ, *which is a solution of* (3).

2. Global existence of strong solution for fixed n.

We aim to solve (4) on the torus $\mathbf{T}^2$ with the additional restriction

$$\sum_{l=1}^n \int_{\mathbf{T}^2} \psi_l = 0.$$

We need this condition in order to ensure the uniqueness of the solution of the elliptic system (4). Indeed, this system corresponds to a discretization of a continuous problem with homogeneous Neuman boundary conditions.

One of the principal goals of this section is to prove the following result.

THEOREM 2.1. *Let* $s > 2$ *and* $(\zeta_1^0, \dots, \zeta_n^0) \in (H^s(\mathbf{T}^2))^n$ *be such that*

$$\sum_{l=1}^n \int_{\mathbf{T}^2} \zeta_l^0 = 0.$$

Then there exists a unique solution $(\zeta_1, \dots \zeta_n) \in (\mathcal{C}(\mathbf{R}^+, H^s(\mathbf{T}^2)))^n$ *and* $((u_1, v_1), \dots, (u_n, v_n)) \in (\mathcal{C}(\mathbf{R}^+, H^{s+1}(\mathbf{T}^2)))^{2n}$ *to* (1) *and* (4) *such that* $(\zeta_1, \dots, \zeta_n)(t = 0) = (\zeta_1^0, \dots, \zeta_n^0)$.

To prove this result, we use a classical energy method to obtain local-in-time existence. Then the solution is shown to extend globally using a priori estimates. The constants occurring in this section may depend on n.

2.1. Local-in-time existence.

PROPOSITION 2.2. *Let* $(\zeta_1^0, \dots, \zeta_n^0) \equiv \zeta^0 \in (H^s(\mathbf{T}^2))^n$ *with* $s > 2$ *such that* $|\zeta^0|_{(H^s)^n} \leq M$. *Then for sufficiently small* T, *there exists a unique solution* $\zeta = (\zeta_1, \dots, \zeta_n)$ *to* (1)–(4) *such that* $\zeta \in \mathcal{C}([0,T], (H^s(\mathbf{T}^2))^n)$ *and* $|\zeta|_{L^\infty([0,T],(H^s(\mathbf{T}^n))^n)} \leq 2M$.

Proof. For $\zeta \in \mathcal{C}([0,T], (H^s(\mathbf{T}^2))^n)$, we construct $\tilde\psi$ satisfying

$$\begin{cases} \zeta_1 = \Delta_{x,y}\tilde\psi_1 + \dfrac{1}{\varepsilon^2}\beta_2(\tilde\psi_2 - \tilde\psi_1), \\ \\ \zeta_l = \Delta_{x,y}\tilde\psi_l - \dfrac{1}{\varepsilon^2}\beta_l(\tilde\psi_l - \tilde\psi_{l-1}) + \dfrac{1}{\varepsilon^2}\beta_{l+1}(\tilde\psi_{l+1} - \tilde\psi_l), \quad 2 \leq l \leq n-1, \\ \\ \zeta_n = \Delta_{x,y}\tilde\psi_n - \dfrac{1}{\varepsilon^2}\beta_n(\tilde\psi_n - \tilde\psi_{n-1}). \end{cases} \tag{16}$$

Let $\tilde{u}_l = -\frac{\partial\tilde{\psi}_l}{\partial y}$ and $\tilde{v}_l = \frac{\partial\tilde{\psi}_l}{\partial x}$ and consider the solution $\tilde{\zeta}$ to

$$\begin{cases} \left(\dfrac{\partial}{\partial t} + \tilde{u}_l \dfrac{\partial}{\partial x} + \tilde{v}_l \dfrac{\partial}{\partial y}\right) \tilde{\zeta}_l = -a\tilde{v}_l, \quad l = 1 \text{ to } n, \\ \tilde{\zeta}(0) = \zeta_0. \end{cases} \tag{17}$$

We denote by $\mathcal{T}$ the mapping that carries ζ into $\tilde{\zeta}$. We want to prove that $\mathcal{T}$ is a contraction in a suitable space $\mathcal{C}(0, T, X)$ for sufficiently small T. Let us first solve the elliptic system. Introduce the operator on $\mathbf{R}^n$ whose matrix is

$$\mathcal{A}_n = -\begin{pmatrix} -\frac{\beta_2}{\varepsilon^2} & \frac{\beta_2}{\varepsilon^2} & & & 0 \\ & \cdots & \cdots & & \\ & \frac{\beta_l}{\varepsilon^2} & -\frac{\beta_l+\beta_{l+1}}{\varepsilon^2} & \frac{\beta_{l+1}}{\varepsilon^2} & \\ & & \cdots & \cdots & \\ 0 & & & \frac{\beta_n}{\varepsilon^2} & -\frac{\beta_n}{\varepsilon^2} \end{pmatrix}.$$

The matrix $\mathcal{A}_n$ is symmetric and

$$(\mathcal{A}_n X, X) = -\sum_{i=1}^{n-1} \beta_{i+1} \left(\frac{x_{i+1} - x_i}{\varepsilon}\right)^2,$$

where $X = (x_1, \ldots, x_n)$, so that—thanks to (3)—0 is a simple eigenvalue. Hence if $\zeta \in (\mathcal{C}([0,T], (H^s)))^n$, then there exists a unique $\tilde{\psi} \in (\mathcal{C}([0,T], (H^{s+2})))^n$ satisfying (16) such that $\sum_{i=1}^n \int_{\mathbf{T}^2} \psi_i dxdy = 0$ and

$$|\tilde{\psi}|_{(\mathcal{C}([0,T],(H^{s+2})))^n} \leq C|\zeta|_{(\mathcal{C}([0,T],(H^s)))^n}. \tag{18}$$

Let us now apply ∂_α^s to (17) and form the product with $\partial_\alpha^s \tilde{\zeta}_l$; an integration yields

$$\frac{\partial}{\partial t}\int_{\mathbf{T}^2} |\partial_\alpha^s \tilde{\zeta}_l|^2 + \int_{\mathbf{T}^2} \partial_\alpha^s((\tilde{u}_l, \tilde{v}_l) \cdot \nabla\tilde{\zeta}_l)\partial_\alpha^s \tilde{\zeta}_l = -a \int_{\mathbf{T}^2} \partial_\alpha^s \tilde{v}_l \partial_\alpha^s \tilde{\zeta}_l. \tag{19}$$

On the other hand, since $\frac{\partial}{\partial x}\tilde{u}_l + \frac{\partial}{\partial y}\tilde{v}_l = 0$, $\int_{\mathbf{T}^2}(\tilde{u}_l, \tilde{v}_l) \cdot \nabla\partial_\alpha^s\tilde{\zeta}_l\partial_\alpha^s\tilde{\zeta}_l = 0$ for all l, and hence (19) becomes

$$\frac{\partial}{\partial t}\int_{\mathbf{T}^2} |\partial_\alpha^s \tilde{\zeta}_l|^2 + \int_{\mathbf{T}^2} (\partial_\alpha^s \nabla((\tilde{u}_l, \tilde{v}_l)\tilde{\zeta}_l) - (\tilde{u}_l, \tilde{v}_l) \cdot \nabla\partial_\alpha^s\tilde{\zeta}_l)\partial_\alpha^s \tilde{\zeta}_l = -a \int_{\mathbf{T}^2} \partial_\alpha^s \tilde{v}_l \partial_\alpha^s \tilde{\zeta}_l. \tag{20}$$

The classical commutator estimate (see [3]) implies

$$\frac{\partial}{\partial t}\int_{\mathbf{T}^2} |\partial_\alpha^s \tilde{\zeta}_l|^2 \leq C(|(\tilde{u}_l, \tilde{v}_l)|_{H^{s+1}}|\tilde{\zeta}_l|_{L^\infty} + |\nabla(\tilde{u}_l, \tilde{v}_l)|_{L^\infty}|\tilde{\zeta}_l|_{H^s})|\tilde{\zeta}_l|_{H^s} + a|\tilde{v}_l|_{H^s}|\tilde{\zeta}_l|_{H^s}. \tag{21}$$

Furthermore, $|\tilde{\zeta}_l|_{L^\infty} \leq C|\tilde{\zeta}_l|_{H^s}$ as soon as $s > 1$ and $|\nabla(\tilde{u}_l, \tilde{v}_l)|_{L^\infty} \leq C|(\tilde{u}_l, \tilde{v}_l)|_{H^{s+1}}$. Hence equation (21) leads to

$$\frac{\partial}{\partial t}\int_{\mathbf{T}^2} |\partial_\alpha^s \tilde{\zeta}_l|^2 \leq C(|(\tilde{u}_l, \tilde{v}_l)|_{H^{s+1}}|\tilde{\zeta}_l|_{H^s}^2) + a|\tilde{v}_l|_{H^s}|\tilde{\zeta}_l|_{H^s}. \tag{22}$$

Thanks to (18), we obtain

$$\frac{\partial}{\partial t}|\tilde{\zeta}_l|^2_{H^s} \leq C_1(1+M)|\tilde{\zeta}_l|^2_{H^s} + C_2$$

if $|\zeta|_{(L^\infty(0,T,H^s))^n} \leq 2M$. This yields

$$\begin{aligned} |\tilde{\zeta}_l|^2_{H^s} &\leq e^{C_1(1+M)t}|\zeta_0|^2_{H^s} + (e^{C_1(1+M)t}-1)\frac{C_2}{C_1(1+M)} \\ &\leq e^{C_1(1+M)t}M + (e^{C_1(1+M)t}-1)\frac{C_2}{C_1(1+M)}. \end{aligned} \tag{23}$$

Choose T sufficiently small so that the right-hand side of (9) is less than or equal to $2M$. If B_M denotes the ball of radius $2M$ in $(\mathcal{C}([0,T],H^s))^n$, then for sufficiently small T,

$$\mathcal{T} \text{ maps } B_M \text{ into itself.} \tag{24}$$

We next show that $\mathcal{T}$ is a contraction in the $(\mathcal{C}([0,T],L^2))^n$ norm, provided that T is sufficiently small. Indeed, for ζ^1 and ζ^2, we have

$$|\tilde{\psi}^1 - \tilde{\psi}^2|_{(L^\infty(0,T,H^{s+2}))^n} \leq C|\zeta^1-\zeta^2|_{(L^\infty(0,T,H^s))^n}. \tag{25}$$

On the other hand,

$$\frac{\partial}{\partial t}(\tilde{\zeta}^1_l - \tilde{\zeta}^2_l) + (\tilde{u}^1_l,\tilde{v}^1_l).\nabla(\tilde{\zeta}^1_l - \tilde{\zeta}^2_l) + ((\tilde{u}^1_l,\tilde{v}^1_l) - (\tilde{u}^2_l,\tilde{v}^2_l))\cdot\nabla\tilde{\zeta}^2_l = -a(\tilde{v}^1_l - \tilde{v}^2_l). \tag{26}$$

Multiply (12) by $(\tilde{\zeta}^1_l - \tilde{\zeta}^2_l)$ and integrate to obtain

$$\begin{aligned} \tfrac{\partial}{\partial t}\textstyle\int_{\mathbf{T}^2}|\tilde{\zeta}^1_l - \tilde{\zeta}^2_l|^2 &\leq C'|\nabla\tilde{\zeta}^2_l|_{L^\infty}|\tilde{\zeta}^1_l - \tilde{\zeta}^2_l|_{L^2}|\zeta^1-\zeta^2|_{(L^2)^n} \\ &\leq C'|\tilde{\zeta}^1_l - \tilde{\zeta}^2_l|_{L^2}|\zeta^1-\zeta^2|_{(L^2)^n} \end{aligned}$$

as soon as $s>2$, where (11) has been used. We deduce that

$$|\tilde{\zeta}^1 - \tilde{\zeta}^2|_{(L^\infty(0,T,L^2))^n} \leq C'T|\zeta^1-\zeta^2|_{(L^\infty(0,T,L^2))^n}.$$

If T is such that $C'T<1$, then $\mathcal{T}$ becomes a contraction from B_R into itself and therefore it has a unique fixed point $\zeta \in (L^\infty(0,T,H^s))^n$. Since $\frac{\partial}{\partial t}\zeta \in (L^\infty(0,T,H^{s-1}))^n$, we get $\zeta \in (\mathcal{C}([0,T],H^{s-\eta}))^n$ for all $\eta>0$. We still have to prove that $\zeta \in (\mathcal{C}([0,T],H^s))^n$. To this end, we use the fact that $(u,v) \in (L^\infty(0,T,H^{s+1}))^n$ and write ζ in term of caracteristics. This concludes the proof of Proposition 2.2.

2.2. Globalization. In order to show that the solution is global, it is sufficient to prove that $|\zeta|_{(H^s)^n}(t)$ cannot tend to infinity in finite time. To do this, we use the fact that (22) can be refined by Youdovitch's techniques [12] for the two-dimensional Euler equation (see also [3] or [8]). Namely, we have $|\zeta(t)|_{(L^\infty)^n} \leq C|\zeta_0|_{(L^\infty)^n}$. It follows that $|(u,v)(t)|_{\mathcal{C}^1_*} \leq C|\zeta_0|_{(L^\infty)^n}$, where $\mathcal{C}^1_*$ denotes Zygmund's class (see [3]). Then for all $\varepsilon>0$, we obtain

$$|\nabla(u,v)(t)|_{(L^\infty)^{2n}} \leq \frac{C}{\varepsilon}|(u,v)|_{\mathcal{C}^1_*}\log\left(e + \frac{|(u,v)|_{\mathcal{C}^{1+\varepsilon}}}{|(u,v)|_{\mathcal{C}^1_*}}\right).$$

(See [3] for a proof of this inequality.) In the present context, this gives

$$\begin{aligned}|\nabla(u,v)(t)|_{(L^\infty)^{2n}} &\le C\log(e+|(u,v)|_{(H^{s+1})^n})\\ &\le C\log(e+|\zeta|_{(H^s)^n})\end{aligned}$$

as soon as $s>2$. Estimate (21) leads to

$$\frac{\partial}{\partial t}|\zeta|^2_{(H^s)^n}\le C|\zeta|^2_{(H^s)^n}\log(e+|\zeta|_{(H^s)^n}),$$

which gives $|\zeta|_{(H^s)^n}\le Ce^{Ce^t}$. Hence the solution is global.

3. Continuous stratification limit. The above results give some bounds on the solution. However, these bounds depend on the number of layers n and therefore are not directly helpful in performing the limit $n\to\infty$. Here we will derive some bounds which are independent of the number of layers and introduce functions depending of a vertical variable z.

Remark 1. The different constants occuring in this section, which we denote generically by C, *do not depend on* n.

In this section, we will use the notation introduced in section 1, especially (6)–(15).

Let $\zeta_0(x,y,z)$ be defined on $\mathbf{T}^2\times[0,1]$, let V_ε be the subspace engendered by $\{(\phi_i^\varepsilon(z)), i=1,\dots,n\}$, and let Π_ε be the projector onto V_ε in L^2. We assume that $\zeta_0(x,y,z)\in L^2\cap\mathcal{C}^0(\mathbf{T}^2\times[0,1])$ and $\zeta_0^\varepsilon=\Pi_\varepsilon\zeta_0$.

The result reads as follows.

THEOREM 3.1. *Let $\zeta_0(x,y,z)\in L^2\cap\mathcal{C}^0(\mathbf{T}^2\times[0,1])$. The following convergences hold for all $0<T<\infty$ when ε tends to* 0:

(i)

$$(U_2^\varepsilon,V_2^\varepsilon)\to(U,V)\quad \textit{in } \mathcal{C}([0,T],L^2(\mathbf{T}^2\times[0,1]))\ \textit{strongly.}$$

(ii) *Z_1^ε and Z_2^ε converge to the same limit Z in $L^p(0,T,L^2(\mathbf{T}^2\times[0,1]))$ strongly for all $p<\infty$ and in $L^\infty(0,T,L^2\cap L^\infty(\mathbf{T}^2\times[0,1]))$ weakly.*

(iii)

$$\Psi_3^\varepsilon\rightharpoonup\Psi\quad \textit{in } L^\infty(0,T,L^2(\mathbf{T}^2\times[0,1]))\ \textit{weakly.}$$

(iv)

$$\begin{aligned}\nabla^\perp\Psi_1^\varepsilon\to\nabla^\perp\Psi\quad &\textit{in } \mathcal{C}([0,T],L^2(\mathbf{T}^2\times[0,1]))\ \textit{strongly}\\ \textit{and}\quad &\textit{in } L^\infty(0,T,H^1(\mathbf{T}^2\times[0,1]))\ \textit{weakly.}\end{aligned}$$

Moreover, (U,V), Ψ, and Z satisfy

$$\frac{\partial Z}{\partial t}+\nabla\cdot((U,V)(Z+ay))=0,$$

$$U=-\frac{\partial\Psi}{\partial y}\quad \textit{and}\quad V=\frac{\partial\Psi}{\partial x},$$

and

$$Z=\Delta_{x,y}\Psi+\frac{\partial}{\partial z}\left(\beta(z)\frac{\partial\Psi}{\partial z}\right),$$

with $\frac{\partial \Psi}{\partial z} = 0$ *in* $z = 0$ *and* $z = 1$,

$$Z(t = 0) = \zeta_0.$$

We have used the classical notation $\nabla^\perp h$ for the vector field $(-\frac{\partial h}{\partial y}, \frac{\partial h}{\partial x})$. The remaining of this section is devoted to the proof of this result.

3.1. A priori estimates. We define the sequences Z_{01}^ε and Z_{02}^ε by (11) and (12). Then Z_{01}^ε and Z_{02}^ε converge to ζ_0 in $L^2(\mathbf{T}^2 \times [0, 1])$ strongly. Moreover, since the functions $\zeta_l^\varepsilon + ay$ are transporting by a measure-preserving flow in time, we have the following:

(27) the sequences Z_1^ε and Z_2^ε are bounded in $L^\infty(0, +\infty, L^2 \cap L^\infty(\mathbf{T}^2 \times [0, 1]))$;

moreover, the following equality holds:

$$|ay + \zeta_l^\varepsilon(x, y, t)|_{L^2 \cap L^\infty(\mathbf{T}^2)} = |ay + \zeta_{0l}^\varepsilon(x, y)|_{L^2 \cap L^\infty(\mathbf{T}^2)}. \tag{28}$$

Definition 3.2. *Let us introduce the space* $H_n^s(\mathbf{T^2})$ *as follows.*

$$H_n^s(\mathbf{T^2}) = \left\{ \psi = (\psi_1, \dots, \psi_n) \in (H^s(\mathbf{T^2}))^n \quad \textit{such that} \int_{\mathbf{T^2}} \sum_{i=1}^n \psi_i = 0 \quad \textit{and} \right.$$
$$\left. \frac{1}{n-1} \sum_{i=1}^n \int_{\mathbf{T^2}} |(-\Delta_{x,y})^{s/2} \psi_i|^2 + \int_{\mathbf{T^2}} \frac{1}{n-1} |(\mathcal{A}_n)^{s/2} \psi|^2 < \infty \right\},$$

where $\mathcal{A}_n$ *is the matrix introduced in section* 2.1.

The space $H_n^s(\mathbf{T^2})$ *is endowed with its natural norm.*

Remark 2. The matrix $\mathcal{A}_n$ corresponds to a discretization of the operator

$$\frac{\partial}{\partial z} \left(\beta(z) \frac{\partial \Psi}{\partial z} \right)$$

with homogeneous Neuman boundary conditions; see, for example, [10].

If we denote the lth eigenvalue of $-\frac{\partial}{\partial z}(\beta(z)\frac{\partial \Psi}{\partial z})$ by λ_l and the lth eigenvalue of $\mathcal{A}_n$ by λ_l^ε, the Min–Max principle (see [5]) implies that $\lambda_l^\varepsilon \geq \lambda_l$. Moreover, $\lambda_0 = 0$ and $\lambda_1 > 0$; hence $\mathcal{A}_n$ has 0 as simple eigenvalue, and the corresponding eigenvector is $(1, \dots, 1)$. Therefore, $((1/(n-1)) \sum_{i=1}^n \int_{\mathbf{T^2}} |(-\Delta_{x,y})^{s/2} \psi_i|^2 + \int_{\mathbf{T^2}} (1/(n-1)) |(\mathcal{A}_n)^{s/2} \psi|^2)^{1/2}$ is a norm on $H_n^s(\mathbf{T^2})$.

Lemma 3.3. *The sequence* (ψ_l^ε) *satisfies*

$$|(\psi_1^\varepsilon, \dots, \psi_n^\varepsilon)|_{L^\infty(\mathbf{R}^+, H_n^2(\mathbf{T^2}))} \leq C,$$

where the constant C *is independent of* n.

Proof. We form the scalar product of (4) with $(\Delta_{x,y} \psi_l^\varepsilon)$ and $(\mathcal{A}_n(\psi_l^\varepsilon))$ and we use the fact that (ζ_l^ε) is bounded in $L^\infty(\mathbf{R}^+, H_n^0(\mathbf{T^2}))$. This yields the result. Moreover, we obtain

$$\frac{1}{n-1} \int_{\mathbf{T^2}} |\nabla_{x,y} \mathcal{A}_n^{1/2}(\psi_l^\varepsilon)|^2 \in L^\infty(\mathbf{R}^+). \tag{29}$$

LEMMA 3.4. *The sequence* (ζ_l^ε) *satisfies*

$$\frac{1}{n-1}\sum_{l=1}^{n}|\frac{\partial\zeta_l^\varepsilon}{\partial t}|_{H^{-1}(\mathbf{T}^2)}\le K$$

and $\frac{\partial\Psi_1^\varepsilon}{\partial t}$ *is bounded in* $L^\infty(\mathbf{R}^+,H^1(\mathbf{T}^2\times[0,1]))$.

Proof. For all l, we have

$$\frac{\partial}{\partial t}\zeta_l^\varepsilon+\nabla_{x,y}\cdot(\nabla_{x,y}^\perp\psi_l^\varepsilon(\zeta_l^\varepsilon+ay))=0.$$

Furthermore,

$$\begin{aligned}|\nabla_{x,y}\cdot(\nabla_{x,y}^\perp\psi_l^\varepsilon(\zeta_l^\varepsilon+ay))|_{H^{-1}(\mathbf{T}^2)}&\le C|\psi_l^\varepsilon|_{H^1(\mathbf{T}^2)}|\zeta_l^\varepsilon|_{L^\infty(\mathbf{T}^2)}\\&\le C|\psi_l^\varepsilon|_{H^1(\mathbf{T}^2)}.\end{aligned}$$

These two last inequalities imply that

$$\left|\frac{\partial\zeta_l^\varepsilon}{\partial t}\right|_{H^{-1}(\mathbf{T}^2)}\le C|\psi_l^\varepsilon|_{H^1(\mathbf{T}^2)}.$$

The first part of the lemma then follows from (29). For the second part, differentiate (4) with respect to t and multiply it by (ψ_l^ε); we obtain that

$$\frac{\partial\psi_l^\varepsilon}{\partial t}\text{ is bounded in }L^\infty(0,T,H_n^1(\mathbf{T}^2)).$$

It then follows that $\frac{\partial\Psi_1^\varepsilon}{\partial t}$ is bounded in $L^\infty(\mathbf{R}^+,H^1(\mathbf{T}^2\times[0,1]))$ by the definition (6) of Ψ_1^ε.

LEMMA 3.5. *The sequences* Ψ_3^ε *and* Z_2^ε *are bounded in* $L^\infty(\mathbf{R}^+,L^2(\mathbf{T}^2\times[0,1]))$.

Proof. It is sufficient to show that the matrix B_n occurring in system (10) is invertible, the norm of its inverse being bounded independently of n. The matrix B_n is clearly symmetric and irreductible. Gerschgörin's theorem implies that its eigenvalues are included in the union of the disks

$$\left|\lambda-\frac{1}{3}\right|\le\frac{1}{6}\quad\text{and}\quad\left|\lambda-\frac{2}{3}\right|\le\frac{1}{3}.$$

The lemma follows.

3.2. End of the proof of Theorem 3.1. Let us consider

$$\nabla_{x,y}^\perp\Psi_1^\varepsilon=\sum_{l=1}^{n}\nabla_{x,y}^\perp\psi_l^\varepsilon(x,y,t)\phi_l^\varepsilon(z).$$

The vector field $\nabla_{x,y}^\perp\Psi_1^\varepsilon$ is bounded in $L^\infty(\mathbf{R}^+,H^1(\mathbf{T}^2\times[0,1]))$ by (29) and $\frac{\partial}{\partial t}\nabla_{x,y}^\perp\Psi_1^\varepsilon$ is bounded in $L^\infty(\mathbf{R}^+,L^2(\mathbf{T}^2\times[0,1]))$ thanks to Lemma 3.4. We can therefore extract a subsequence that converges in $\mathcal{C}([0,T],L^2(\mathbf{T}^2\times[0,1]))$ strongly and in $L^\infty(0,T,H^1(\mathbf{T}^2\times[0,1]))$ weakly for all $T<\infty$ to a vector field $\nabla_{x,y}^\perp\Psi$. Moreover, $Z_1^\varepsilon\rightharpoonup Z$ in $L^\infty(0,T,L^2\cap L^\infty)$ weakly. We will now show that the limits of the sequences Z_1^ε and Z_2^ε are the same and that this is also the case for the limits of Ψ_1^ε, Ψ_2^ε, and Ψ_3^ε.

The function Ψ_1^ε is related to a P1 finite-element approximation in the z-direction of Ψ, while Ψ_2^ε is a piecewise-constant approximation of Ψ_1^ε. It will therefore be possible to show that $\nabla_{x,y}^\perp \Psi_2^\varepsilon - \nabla_{x,y}^\perp \Psi_1^\varepsilon \rightarrow_{\varepsilon\rightarrow 0} 0$ in $L^\infty([0,T], L^2(\mathbf{T}^2\times[0,1]))$ strongly; this is related to the fact that $\nabla_{x,y}^\perp \Psi_1^\varepsilon$ is bounded in $L^\infty([0,T], H^1(\mathbf{T}^2\times[0,1]))$.

The functions Ψ_3^ε and Z_2^ε are also some piecewise-constant approximations of Ψ and Z, but these approximations are obtained versus the mass matrix B_n, and we cannot directly use the H^1 bound of Ψ_1^ε. Therefore, the convergences $\Psi_3^\varepsilon - \Psi \rightarrow 0$ and $Z_2^\varepsilon - Z \rightarrow 0$ are obtained only in $L^\infty([0,T], L^2(\mathbf{T}^2\times[0,1]))$ weakly.

These results are stated precisely in the next two propositions and proved by some explicit computations.

PROPOSITION 3.6. *The sequence $\nabla_{x,y}^\perp \Psi_2^\varepsilon$ converges to $\nabla_{x,y}^\perp \Psi$ in $\mathcal{C}([0,T], L^2(\mathbf{T}^2\times[0,1]))$ strongly.*

Proof. We compute the $L^2(\mathbf{T}^2\times[0,1])$ norm of the difference between $\nabla_{x,y}^\perp \Psi_2^\varepsilon$ and $\nabla_{x,y}^\perp \Psi_1^\varepsilon$ and get

$$\begin{aligned}
&|\nabla_{x,y}^\perp \Psi_2^\varepsilon - \nabla_{x,y}^\perp \Psi_1^\varepsilon|_{L^2(\mathbf{T}^2\times[0,1])} \\
&\quad = \int_{\mathbf{T}^2}\int_0^1 \left| \sum_{l=1}^{n} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\phi_l^\varepsilon(z) - \sum_{l=1}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\mathbf{1}_{[(l-1)\varepsilon,l\varepsilon[}(z)\right|^2 dxdydz.
\end{aligned}$$

Taking into account the supports of ϕ_l^ε, we obtain

$$\begin{aligned}
&\int_0^1 \left| \sum_{l=1}^{n} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\phi_l^\varepsilon(z) - \sum_{l=1}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\mathbf{1}_{[(l-1)\varepsilon,l\varepsilon[}(z)\right|^2 dz \\
&\quad = \int_0^1 \sum_{l=1}^{n} |\nabla_{x,y}^\perp \psi_l^\varepsilon|^2 |\phi_l^\varepsilon|^2 dz + \frac{1}{n-1}\sum_{l=1}^{n-1} |\nabla_{x,y}^\perp \psi_l^\varepsilon|^2 \\
&\quad - 2\int_0^1 \sum_{l=1}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon \phi_l^\varepsilon \nabla_{x,y}^\perp \psi_l^\varepsilon \mathbf{1}_{[(l-1)\varepsilon,l\varepsilon[} \\
&\quad - 2\int_0^1 \sum_{l=2}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon \phi_l^\varepsilon \nabla_{x,y}^\perp \psi_{l-1}^\varepsilon \mathbf{1}_{[(l-2)\varepsilon,(l-1)\varepsilon[} \\
&\quad + 2\int_0^1 \sum_{l=1}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon \phi_l^\varepsilon \nabla_{x,y}^\perp \psi_{l+1}^\varepsilon \phi_{l+1}^\varepsilon.
\end{aligned}$$

An explicit computation of the integrals yields

$$\begin{aligned}
&\int_0^1 \left| \sum_{l=1}^{n} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\phi_l^\varepsilon(z) - \sum_{l=1}^{n-1} \nabla_{x,y}^\perp \psi_l^\varepsilon(x,y,t)\mathbf{1}_{[(l-1)\varepsilon,l\varepsilon[}(z)\right|^2 dz \\
&\quad = \frac{2\varepsilon}{3}\sum_{l=2}^{n-2} \nabla_{x,y}^\perp \psi_l^\varepsilon \cdot (\nabla_{x,y}^\perp \psi_l^\varepsilon - \nabla_{x,y}^\perp \psi_{l-1}^\varepsilon) \qquad (30) \\
&\quad - \frac{2\varepsilon}{3}\nabla_{x,y}^\perp \psi_n^\varepsilon \cdot \nabla_{x,y}^\perp \psi_{n-1}^\varepsilon + \frac{\varepsilon}{3}|\nabla_{x,y}^\perp \psi_1^\varepsilon|^2 + \frac{\varepsilon}{3}|\nabla_{x,y}^\perp \psi_n^\varepsilon|^2.
\end{aligned}$$

Then since $(\nabla^{\perp}_{x,y}\psi^{\varepsilon}_1)_l$ is bounded in $L^{\infty}(\mathbf{R}^+, H^1_n(\mathbf{T}^2))$, we have

$$\left|\int_{\mathbf{T}^2} \frac{2\varepsilon}{3}\sum_{l=2}^{n-2} \nabla^{\perp}_{x,y}\psi^{\varepsilon}_l . (\nabla^{\perp}_{x,y}\psi^{\varepsilon}_l - \nabla^{\perp}_{x,y}\psi^{\varepsilon}_{l-1})\right| \le C\frac{2\varepsilon}{3} \underset{\varepsilon\to 0}{\longrightarrow} 0$$

in $L^{\infty}(0,T)$. We still have to deal with the following terms in (30):

$$-\frac{2\varepsilon}{3}\nabla^{\perp}_{x,y}\psi^{\varepsilon}_n . \nabla^{\perp}_{x,y}\psi^{\varepsilon}_{n-1} + \frac{\varepsilon}{3}|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_1|^2 + \frac{\varepsilon}{3}|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_n|^2.$$

Since $\nabla^{\perp}_{x,y}\Psi^{\varepsilon}_1$ is bounded in $L^{\infty}(\mathbf{R}^+, H^1(\mathbf{T}^2\times[0,1]))$, for almost every x and y and for all i, we have

$$|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_i|^2_{L^{\infty}(0,1)} \le C|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_i|^2_{H^1(0,1)}.$$

Hence

$$\int_{\mathbf{T}^2} |\nabla^{\perp}_{x,y}\psi^{\varepsilon}_i|^2 \le C|\nabla^{\perp}_{x,y}\Psi^{\varepsilon}_1|^2_{H^1((0,1)\times\mathbf{T}^2)}$$

and

$$\int_{\mathbf{T}^2}\left(-\frac{2\varepsilon}{3}\nabla^{\perp}_{x,y}\psi^{\varepsilon}_n . \nabla^{\perp}_{x,y}\psi^{\varepsilon}_{n-1} + \frac{\varepsilon}{3}|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_1|^2 + \frac{\varepsilon}{3}|\nabla^{\perp}_{x,y}\psi^{\varepsilon}_n|^2\right) \underset{\varepsilon\to 0}{\longrightarrow} 0$$

in $L^{\infty}(0,T)$. Plugging this result into (30) and integrating on $\mathbf{T}^2$ conclude the proof of the proposition.

Concerning the sequences Ψ^{ε}_3 and Z^{ε}_2, we have the following result.

PROPOSITION 3.7. *The sequences Ψ^{ε}_3 and Z^{ε}_2 converge, respectively, to Ψ and Z in $L^{\infty}(0,T,L^2(\mathbf{T}^2\times[0,1]))$ weakly.*

Proof. Since the functions Ψ^{ε}_3 and Z^{ε}_2 are obtained by the same construction, it is enough to show the result for one of them. We will work with Z^{ε}_2 and show that $Z^{\varepsilon}_1 - Z^{\varepsilon}_2 \rightharpoonup 0$ in $L^{\infty}(0,T,L^2(\mathbf{T}^2\times[0,1]))$.

For any interval $[a,b]\subset[0,1]$, for fixed n, there exist two integers k_1 and k_2 such that $(k_1-1)/(n-1) < a \le k_1/(n-1)$ and $k_2/(n-1) \le b < (k_2+1)/(n-1)$. Then

$$\int_a^b (Z^{\varepsilon}_1 - Z^{\varepsilon}_2)dz = \int_a^{\frac{k_1}{n-1}} (Z^{\varepsilon}_1 - Z^{\varepsilon}_2)dz + \int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} (Z^{\varepsilon}_1 - Z^{\varepsilon}_2)dz + \int_{\frac{k_2}{n-1}}^{b} (Z^{\varepsilon}_1 - Z^{\varepsilon}_2)dz. \tag{31}$$

Let us compute each term of the right-hand side of (31).

(i)

$$\begin{aligned}\int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} (Z^{\varepsilon}_1 - Z^{\varepsilon}_2)dz &= \int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} \sum_{l=k_1+1}^{k_2} \zeta^{\varepsilon}_l \mathbf{1}_{[(l-1)\varepsilon, l\varepsilon[} dz - \int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} \sum_{l=k_1-1}^{k_2} g^{\varepsilon}_l \phi^{\varepsilon}_l dz \\ &= \varepsilon\sum_{l=k_1+1}^{k_2}\zeta^{\varepsilon}_l - g^{\varepsilon}_{k_1-1}\frac{\varepsilon}{2} - g^{\varepsilon}_{k_2}\frac{\varepsilon}{2} - \varepsilon\sum_{l=k_1}^{k_2-1} g^{\varepsilon}_l .\end{aligned}$$

Using the relationship between ζ^{ε}_l and g^{ε}_l, namely $\zeta^{\varepsilon}_l = (1/6)(g^{\varepsilon}_{l-1} + 4g^{\varepsilon}_l + g^{\varepsilon}_{l+1})$ for $1 < l < n$, we get after simplification that

$$\int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} (Z_1^\varepsilon - Z_2^\varepsilon)dz = \varepsilon\zeta_{k_2}^\varepsilon - \frac{\varepsilon}{3}g_{k_2}^\varepsilon - \frac{\varepsilon}{3}g_{k_2-1}^\varepsilon - \frac{\varepsilon}{2}g_{k_1-1}^\varepsilon - \frac{\varepsilon}{3}g_{k_1+1}^\varepsilon - \frac{5\varepsilon}{6}g_{k_1}^\varepsilon. \tag{32}$$

Therefore, equation (32) yields

$$\left|\int_{\frac{k_1}{n-1}}^{\frac{k_2}{n-1}} (Z_1^\varepsilon - Z_2^\varepsilon)dz\right| \leq C\sqrt{\varepsilon}\left(\varepsilon\sum_{l=1}^n g_l^{\varepsilon 2}\right)^{1/2}; \tag{33}$$

the constant C does not depend on n or $[a,b]$.

(ii) We proceed in the same way for the terms $\int_a^{k_1/(n-1)}$ and $\int_{k_2/(n-1)}^b$, and as in (33), we finally obtain

$$\left|\int_a^b (Z_1^\varepsilon - Z_2^\varepsilon)dz\right| \leq C\sqrt{\varepsilon}\left(\varepsilon\sum_{l=1}^n g_l^{\varepsilon 2}\right)^{1/2}. \tag{34}$$

Now take $\phi \in L^1(0,T,L^2(\mathbf{T}^2 \times [0,1]))$ and construct

$$\phi^N = \sum_{i=1}^N \alpha_i^N(x,y,t)\mathbf{1}_{[a_i(x,y,t),b_i(x,y,t)[}(z)$$

such that

$$|\phi - \phi^N|_{L^1(0,T,L^2(\mathbf{T}^2\times[0,1]))} \underset{N\to\infty}{\longrightarrow} 0.$$

Then

$$\begin{aligned}&\int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi dxdydzdt \\ &\quad= \int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)(\phi - \phi^N) + \int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi^N.\end{aligned}$$

The first estimate that we have is

$$\left|\int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi\right| \leq C|\phi - \phi^N|_{L^1(0,T,L^2(\mathbf{T}^2\times[0,1]))} + \left|\int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi^N\right|.$$

Hence by (24),

$$\begin{aligned}\left|\int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi^N\right| &\leq \int_0^T \int_{\mathbf{T}^2} C\sqrt{\varepsilon}\left(\sum_{i=1}^N |\alpha_i^N(x,y,t)|\right)\left(\varepsilon\sum_{l=1}^n g_l^{\varepsilon 2}\right)^{1/2} \\ &\leq \sqrt{\varepsilon}\left|\left(\sum_{i=1}^N |\alpha_i^N(x,y,t)|\right)\right|_{L^2((0,T)\times\mathbf{T}^2)}\end{aligned}$$

since $(\varepsilon\sum_{l=1}^n g_l^{\varepsilon 2})^{1/2}$ is bounded in $L^\infty((0,T),L^2(\mathbf{T}^2))$. We therefore obtain

$$\left|\int_0^T \int_{\mathbf{T}^2}\int_0^1 (Z_1^\varepsilon - Z_2^\varepsilon)\phi\right| \leq C|\phi - \phi^N|_{L^1(0,T,L^2(\mathbf{T}^2\times[0,1]))}$$

$$+ C\sqrt{\varepsilon}\left|\left(\sum_{i=1}^{N}|\alpha_i^N(x,y,t)|\right)\right|_{L^2((0,T)\times\mathbf{T}^2)}.$$

It follows that

$$\limsup_{\varepsilon\to 0}\left|\int_0^T\int_{\mathbf{T}^2}\int_0^1(Z_1^\varepsilon - Z_2^\varepsilon)\phi\right| \le C|\phi-\phi^N|_{L^1(0,T,L^2(\mathbf{T}^2\times[0,1]))}.$$

Letting $N\to\infty$ in this expression, we obtain the result.

We can now perform the limit process on (13), (14), and (15). Thanks to Proposition 3.6, (15) directly gives

$$U = -\frac{\partial\Psi}{\partial y} \quad\text{and}\quad V = \frac{\partial\Psi}{\partial x}.$$

On the other hand, since $Z_1^\varepsilon \rightharpoonup Z$ in $L^\infty(0,T,L^2\cap L^\infty)$ weakly and since $(U_2^\varepsilon, V_2^\varepsilon)\to(U,V)$ in $L^\infty(0,T,L^2)$ strongly, we have

$$(U_2^\varepsilon,V_2^\varepsilon)(Z_1^\varepsilon + ay) \rightharpoonup (U,V)(Z+ay) \quad \text{in } \mathcal{D}',$$

and (13) gives

$$\frac{\partial Z}{\partial t} + \nabla\cdot((U,V)(Z+ay)) = 0.$$

In order to perform the limit on the elliptic equation (14), one use Proposition 3.7 and the classical result of variational approximation.

Moreover, since the L^2 norm of $Z+ay$ is conserved by the flow and since $Z_{01}^\varepsilon\to\zeta_0$ in L^2 strongly, for all t the convergences of $Z_1^\varepsilon(t)$ and $Z_2^\varepsilon(t)$ are strong in L^2; hence $Z_1^\varepsilon(t)$ and $Z_2^\varepsilon(t)$ converge to $Z(t)$ in $L^p(0,T,L^2(\mathbf{T}^2\times[0,1]))$ strongly for all $p<\infty$.

REFERENCES

[1] A. J. Bourgeois and J. T. Beale, *Validity of the quasigeostrophic model for large scale flow in the atmosphere and ocean,* SIAM J. Math. Anal., 25 (1994), pp. 1023–1068.

[2] J. Y. Chemin, *A propos d'un problème de pénalisation de type anti-symétrique,* C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 861–864.

[3] J. Y. Chemin, *Fluides parfaits incompressibles,* SMF Collection, Astérisque, 230 (1995).

[4] T. Colin, *Remarks on an homogenous model of ocean circulation,* Asymptotic Anal., 12 (1996), pp. 153–168.

[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics,* Vol. 1, John Wiley, New York, 1989.

[6] G. Evensen, *Inverse methods and data assimilation in nonlinear ocean models,* Phys. D, 77 (1994), pp. 108–129.

[7] E. Grenier, *Fluides en rotations et ondes d'inertie*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 711–714.

[8] T. Kato and G. Ponce, *Commutator estimates and the Euler and Navier–Stokes equations*, Comm. Pure Appl. Math., XLI (1988), pp. 891–907.

[9] J. K. Pedlosky, *Geophysical Fluid Dynamics,* 2nd ed., Springer-Verlag, Berlin, 1987.

[10] P. A. Raviart and J. M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles,* Masson, Paris, 1988.

[11] A. Wiin-Nielsen, *Nonlinear studies of quasi-geostrophic systems,* Phys. D, 77 (1994), pp. 33–59.

[12] V. I. Youdovitch, *Flot instationnaire d'un liquide idéal incompressible,* Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), pp. 1032–1066.

SIAM J. Math. Anal.
Vol. 28, No. 3, pp. 530–538, May 1997

HIGHER GRADIENT INTEGRABILITY OF MINIMIZERS FOR A POLYCONVEX CASE IN TWO DIMENSIONS[*]

MICHAEL M. DOUGHERTY[†]

Abstract. Local L^p-estimates, $1 < p < \infty$, of the gradient are proved for minimizers of certain functionals whose integrands have quadratic growth in the gradient and are polyconvex "at infinity." Both variations of dependent and independent variables of the minimizers are used to derive equations from which Calderon–Zygmund theory gives the result. Hölder continuity of the minimizers will follow as a corollary.

Key words. polyconvexity, regularity theory

AMS subject classifications. Primary, 49N60; Secondary, 35J60, 73C50

PII. S0036141095292585

1. Introduction.

In this paper, local higher integrability of the gradient is proved for minimizers of certain polyconvex functionals in the calculus of variations. Specifically, let $\Omega \subset \mathbb{R}^2$ be a bounded domain. Define an energy functional $\mathcal{E}$ by

$$\mathcal{E}[\mathcal{U}] = \int_\Omega \left(\frac{1}{2} |D\mathcal{U}|^2 + h(\det D\mathcal{U}) \right) dX = \int_\Omega \gamma(D\mathcal{U})\, dX \tag{1.1}$$

for each $\mathcal{U} = (u, v) \in W^{1,2}(\Omega; \mathbb{R}^2)$, where $h \in C^1(\mathbb{R})$ is convex, $h \geq 0$, and $|h'|$ is bounded. Thus

$$\lim_{d\to\infty} h'(d) = L_+ \quad \text{and} \quad \lim_{d\to-\infty} h'(d) = -L_- \tag{1.2}$$

exist and are finite. Also assume that h is asymptotically linear in the sense that there exists $\beta > 0$ such that

$$\begin{aligned} L_+ - h'(d) &= \mathcal{O}(d^{-\beta}) \quad &&\text{as } d \to \infty, \\ -L_- - h'(d) &= \mathcal{O}(|d|^{-\beta}) \quad &&\text{as } d \to -\infty. \end{aligned} \tag{1.3}$$

The function $\mathcal{U}$ is called a *minimizer* of $\mathcal{E}$ if $\mathcal{E}(\mathcal{U}) \leq \mathcal{E}(\mathcal{W})$ whenever $\mathcal{U} - \mathcal{W} \in W_0^{1,2}(\Omega; \mathbb{R}^2)$. The main result is the following.

Theorem. *Let $\mathcal{U} \in W^{1,2}(\Omega; \mathbb{R}^2)$ be a minimizer for the functional $\mathcal{E}$ given by* (1.1) *and* (1.3)*, with $\beta > 0$. Then $|D\mathcal{U}| \in L^p_{\text{loc}}(\Omega)$ for any $p \in [1, \infty)$. Moreover, for each $p \in [1, \infty)$ and $\Omega' \subset\subset \Omega$,*

$$\|D\mathcal{U}\|_{L^p(\Omega')} \leq C\big(1 + \|\mathcal{U}\|_{L^2(\Omega)} + \|D\mathcal{U}\|_{L^2(\Omega)}\big), \tag{1.4}$$

where $C < \infty$ depends on p, Ω', Ω, and the structure of h.

The Sobolev and Morrey embedding theorems then give the following.

Corollary. *If $\mathcal{U}$ is as in the theorem, then $\mathcal{U} \in C^{0,\alpha}(\Omega; \mathbb{R}^2)$ for any $\alpha \in (0, 1)$.*

Prior work on higher integrability includes a result of Giaquinta and Giusti (see [7, Theorem 4.1] or [6, Theorem 3.1, p. 159]). Applying their work to minimizers of (1.1)

[*] Received by the editors September 29, 1995; accepted for publication February 7, 1996. This research was partially supported by NSF grant DMS-9306199 and is a portion of the author's Ph.D. thesis, supervised by Daniel Phillips.

http://www.siam.org/journals/sima/28-3/29258.html

[†] Department of Mathematics, Penn State University, Berks Campus, Reading, PA 19610 (mmd@math.psu.edu)

gives $|D\mathcal{U}| \in L^{2+\sigma}_{\mathrm{loc}}(\Omega)$ for some $\sigma > 0$. Their argument is based on the minimality of $\mathcal{U}$ and polynomial growth in $D\mathcal{U}$ of the integrand, but it relies neither on any convexity nor on differentiability of the integrand. A result under more stringent conditions is given by Chipot and Evans [3], who proved local Lipschitz regularity for a class of minimizers $\mathcal{U} \in W^{1,2}(\Omega; \mathbb{R}^N)$, $\Omega \subset \mathbb{R}^n$, of functionals whose Euler–Lagrange equations become linear, strongly elliptic constant-coefficient equations as $|D\mathcal{U}| \to \infty$. Their estimates arise through a blowup method. For n-dimensional cases, $n \geq 2$, where $\gamma(D\mathcal{U}) = |D\mathcal{U}|^p + h(\det D\mathcal{U})$ and $|h(\det D\mathcal{U})| \leq A + B|D\mathcal{U}|^q$, $1 \leq q < p < \infty$, Dougherty and Phillips [5] proved (1.4) by applying a result of DiBenedetto and Manfredi [4] on solutions of nonhomogeneous p-Laplace equations.

In this paper, the functional (1.1) is differentiable, but the integrand remains polyconvex at infinity. Indeed, though $h(d)$ becomes linear for large d, this term can have large oscillations for large $|D\mathcal{U}|$. The method below exploits two systems of equations derived from (1.1), these being the usual weak Euler–Lagrange equations and another system derived from variations of the independent variables. The L^p-boundedness of Riesz potentials (see [8] and [4]) yields the main result.

2. Preliminaries.

LEMMA 2.1. *Minimizers of* (1.1) *exist.*

Proof. This follows from work of Ball and Murat, who showed that $\mathcal{E}$ is sequentially weakly lower semicontinuous and thus that the integrand in (1.1) is $W^{1,2}$-quasiconvex [1, Theorem 4.5]. □

The weak Euler–Lagrange equations can be obtained from a straightforward application of the dominated-convergence theorem. Specifically, it is not hard to show that the following lemma holds (cf. [6, Chapter I]).

LEMMA 2.2. *Let $\mathcal{E}$ be as in* (1.1). *If $\mathcal{U}$ is a minimizer of $\mathcal{E}$, then the Euler–Lagrange equations hold in the weak sense:*

$$\Delta u = \bigtriangledown \cdot \big\langle -h'(d)v_y,\, h'(d)v_x \big\rangle, \tag{2.1.1}$$

$$\Delta v = \bigtriangledown \cdot \big\langle h'(d)u_y,\, -h'(d)u_x \big\rangle, \tag{2.1.2}$$

where $d = \det D\mathcal{U} = u_x v_y - u_y v_x$.

A simple density argument shows that $h(d)$ can be replaced by any function

$$\hat{h}(d) = h(d) + a \cdot d + b, \tag{2.2}$$

where a and b are arbitrary constants, without changing the set of minimizers. Indeed, it is not hard to prove the following.

LEMMA 2.3. *For any $\mathcal{U} \in W^{1,2}(\Omega; \mathbb{R}^2)$ and any $\Phi \in W^{1,2}_0(\Omega; \mathbb{R}^2)$,*

$$\int_\Omega \big(\det(D\mathcal{U} + D\Phi) - \det(D\mathcal{U})\big)\, dX = 0. \tag{2.3}$$

Choosing $a = (L_- - L_+)/2$ and adjusting b so $h \geq 0$ still holds, it can be assumed without loss of generality that $L_- = L_+ = M > 0$. Equation (1.3) can then be replaced by

$$\begin{aligned} |M - h'(d)| &\leq K|d|^{-\beta} \quad \text{for } d \geq 1, \\ |-M - h'(d)| &\leq K|d|^{-\beta} \quad \text{for } d \leq -1. \end{aligned} \tag{2.4}$$

From now on, it will also be assumed without loss of generality that $\beta \in (0, 1)$.

Next, consider equilibrium equations which arise from variations of the independent variables. For each $\Phi \in C_c^1(\Omega;\mathbb{R}^2)$, define $\mathcal{Z}_\varepsilon(X) \equiv X + \varepsilon\Phi(X)$. For $|\varepsilon|$ sufficiently small, $\mathcal{U}(\mathcal{Z}_\varepsilon) - \mathcal{U} \in W_0^{1,2}(\Omega;\mathbb{R}^2)$, and so $\mathcal{E}[\mathcal{U}] \le \mathcal{E}[\mathcal{U}(\mathcal{Z}_\varepsilon)]$. An argument similar to that given in [2, Theorem A.1] allows differentiation of $\mathcal{E}[\mathcal{U}(\mathcal{Z}_\varepsilon)]$ in ε to get the following.

Lemma 2.4. *For $\mathcal{U}$ and $\mathcal{E}$ as in the theorem, the following holds for every $\Phi \in C_c^1(\Omega;\mathbb{R}^2)$:*

$$0 = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \mathcal{E}[\mathcal{U}(\mathcal{Z}_\varepsilon)] = \int_\Omega \left(\frac{\partial\gamma}{\partial \mathcal{U}^i_{x_j}}(D\mathcal{U})\mathcal{U}^i_{x_k}\Phi^k_{x_j} - \gamma(D\mathcal{U})\left(\Phi^1_{x_1} + \Phi^2_{x_2}\right)\right) dX. \tag{2.5}$$

A straightforward calculation then shows this system can be written in $\mathcal{D}'(\Omega)$ as

$$\left[\frac{u_x^2 + v_x^2 - u_y^2 - v_y^2}{2} + f(d)\right]_x + [u_x u_y + v_x v_y]_y = 0, \tag{2.6.1}$$

$$[u_x u_y + v_x v_y]_x + \left[\frac{u_y^2 + v_y^2 - u_x^2 - v_x^2}{2} + f(d)\right]_y = 0, \tag{2.6.2}$$

where

$$f(d) = dh'(d) - h(d). \tag{2.7}$$

Lemma 2.5. *For $\mathcal{E}$ and $\mathcal{U}$ as in Lemma 1.1, $f \in L^{\frac{1}{1-\beta}}(\Omega)$.*

Proof. This follows from the asymptotic behavior of h. For $d \ge 1$,

$$\begin{aligned} |f(d)| &= \left|d(h'(d) - M) + dM - h(d)\right| \\ &\le \left|d(h'(d) - M)\right| + \left|\int_0^d \left(M - h'(s)\right) ds - h(0)\right| \\ &\le Kd^{1-\beta} + 2M + \int_1^d Ks^{-\beta}\, ds + |h(0)|. \end{aligned}$$

Replacing M with $-M$ gives a similar result for $d \le -1$, and so for all d,

$$|f(d)| \le C(1 + |d|)^{1-\beta}, \tag{2.8}$$

which gives

$$\|f(d)\|_{L^{\frac{1}{1-\beta}}(\Omega)} \le C\left(1 + \|D\mathcal{U}\|^{2(1-\beta)}_{L^2(\Omega)}\right) \le C\left(1 + \|D\mathcal{U}\|^2_{L^2(\Omega)}\right). \qquad \square \tag{2.9}$$

Now define the following quantities:

$$\begin{aligned} A &= \frac{u_x^2 + v_x^2 - u_y^2 - v_y^2}{2}, \\ B &= u_x u_y + v_x v_y. \end{aligned} \tag{2.10}$$

Subtracting $\partial/\partial y$ of (2.6.2) from $\partial/\partial x$ of (2.6.1) and adding $\partial/\partial x$ of (2.6.2) to $\partial/\partial y$ of (2.6.1) gives in $\mathcal{D}'(\Omega)$ the system

$$\begin{aligned} \Delta A &= -f_{xx} + f_{yy}, \\ \Delta B &= -2f_{xy}. \end{aligned} \tag{2.11}$$

The proof of the following lemma follows an argument similar to that given in [2, pp. 126–127].

LEMMA 2.6. *For $\mathcal{U}$ and $\mathcal{E}$ as in the theorem and $p \in (1,\infty)$, if $f(d) \in L^p(B_{3r})$, $B_{3r} \subset \Omega$, then $A, B \in L^p(B_r)$, where $B_r \subset B_{3r}$ are concentric. Furthermore,*

$$\|A\|_{L^p(B_r)} + \|B\|_{L^p(B_r)} \le c_{r,p} \Big(\|f\|_{L^p(B_{3r})} + \|D\mathcal{U}\|^2_{L^2(B_{3r})} \Big) . \tag{2.12}$$

Proof. Let $\eta \in C_c^\infty(B_{3r})$, with $0 \le \eta \le 1$, $\eta = 1$ on B_{2r}, $|\triangledown\eta| \le cr^{-1}$, and $|D^2\eta| \le cr^{-2}$, with c independent of r. Let $g \in C_c^\infty(B_r)$. Set

$$w(X) = \frac{1}{2\pi} \int_{B_r} \log |X - Z| g(Z)\, dZ. \tag{2.13}$$

Then

$$\begin{aligned}
\int_{B_r} Ag\, dX &= \int_{\mathbb{R}^2} A\eta\Delta w\, dX = \int_{\mathbb{R}^2} \big(A\left(\Delta(\eta w) - w\Delta\eta - 2\triangledown w\cdot\triangledown\eta\right)\big)\, dX \\
&= \int_{\mathbb{R}^2} f(d)\big(-(\eta w)_{xx} + (\eta w)_{yy}\big)\, dX + \int_{B_{3r}-B_{2r}} A\left(-w\Delta\eta - 2\triangledown w\cdot\triangledown\eta\right)\, dX \\
&= \int_{B_{3r}} f(d)\eta(-w_{xx} + w_{yy})\, dX \\
+ &\int_{B_{3r}-B_{2r}} \Big\{ f(d)(-2\eta_x w_x - \eta_{xx} w + 2\eta_y w_y + \eta_{yy} w) - Aw\Delta\eta - 2A\triangledown w\cdot\triangledown\eta \Big\}\, dX \\
&\equiv \text{(I)} + \text{(II)}.
\end{aligned}$$

Now let $1 < p < \infty$ and $q = p/(p-1)$. By the Hölder and Calderon–Zygmund inequalities, it follows that

$$\begin{aligned}
|\text{I}| &\le c_p \|f\|_{L^p(B_{3r})} \|D^2 w\|_{L^q(B_r)} \le c_p \|f\|_{L^p(B_{3r})} \|D^2 w\|_{L^q(\mathbb{R}^2)} \\
&\le c_p' \|f\|_{L^p(B_{3r})} \|g\|_{L^q(\mathbb{R}^2)} = c_p' \|f\|_{L^p(B_{3r})} \|g\|_{L^q(B_r)}.
\end{aligned} \tag{2.14}$$

To estimate (II), note first that Hölder's inequality gives

$$|w| + |\triangledown w| \le c(r,p)\, \|g\|_{L^q(B_r)} \quad \text{on } B_{3r} - B_{2r}.$$

Thus

$$\begin{aligned}
\text{(II)} &\le c'(r,p) \big\{ \|f\|_{L^1(B_{3r})} + \|A\|_{L^1(B_{3r})} \big\}\, \|g\|_{L^q(B_r)} \\
&\le c'(r,p) \big\{ \|f\|_{L^1(B_{3r})} + \|D\mathcal{U}\|^2_{L^2(B_{3r})} \big\}\, \|g\|_{L^q(B_r)}.
\end{aligned} \tag{2.15}$$

Combining (2.14) and (2.15) and taking the supremum over all $g \in C_c^\infty(B_r)$ with $\|g\|_{L^q} = 1$ gives

$$\begin{aligned}
\|A\|_{L^p(B_r)} &\le c''(r,p) \big\{ \|f\|_{L^1(B_{3r})} + \|f\|_{L^p(B_r)} + \|D\mathcal{U}\|^2_{L^2(B_{3r})} \big\} \\
&\le c'''(r,p) \big\{ \|f\|_{L^p(B_{3r})} + \|D\mathcal{U}\|^2_{L^2(B_{3r})} \big\}.
\end{aligned}$$

This proves estimate (2.12) for A, and the estimate for B follows similarly. □

Now consider the quantities $\nu_1(x), \nu_2(x) \ge 0$, defined to be the singular values of $D\mathcal{U}$, i.e., the eigenvalues of $\sqrt{D\mathcal{U}^T D\mathcal{U}}$. Thus ν_1^2 and ν_2^2 will each satisfy the characteristic equation for the matrix

$$D\mathcal{U}^T D\mathcal{U} = \begin{pmatrix} u_x^2 + v_x^2 & u_x u_y + v_x v_y \\ u_x u_y + v_x v_y & u_y^2 + v_y^2 \end{pmatrix} \tag{2.16}$$

given by

$$0 = \det \begin{pmatrix} (u_x^2+v_x^2)-\lambda & u_xu_y+v_xv_y \\ u_xu_y+v_xv_y & (u_y^2+v_y^2)-\lambda \end{pmatrix} = \lambda^2 - |D\mathcal{U}|^2\lambda + d^2. \tag{2.17}$$

Hence

$$\left(\nu_1^2-\nu_2^2\right)^2 = |D\mathcal{U}|^4 - 4d^2 = 4\left(\frac{|D\mathcal{U}|^2}{2}+|d|\right)\left(\frac{|D\mathcal{U}|^2}{2}-|d|\right), \tag{2.18}$$

$$\nu_1^2+\nu_2^2 = |D\mathcal{U}|^2. \tag{2.19}$$

This leads to the following.

Lemma 2.7. *For $\mathcal{E}$ and $\mathcal{U}$ as in Lemma* 1.1 *and for $p \in (1,\infty)$, if $f(d) \in L^p(B_{3r})$, then $\nu_1^2-\nu_2^2 \in L^p(B_r)$. In particular, the following estimates hold:*

$$\|\nu_1-\nu_2\|_{L^{2p}(B_r)} \le C\left(1+\|f\|_{L^p(B_{3r})}+\|D\mathcal{U}\|_{L^2(B_{3r})}\right); \tag{2.20}$$

$$\left\|\sqrt{\frac{|D\mathcal{U}|^2}{2}-|d|}\right\|_{L^{2p}(B_r)} \le C\left(1+\|f\|_{L^p(B_{3r})}+\|D\mathcal{U}\|_{L^2(B_{3r})}\right). \tag{2.21}$$

Proof. For almost every point $X \in \Omega$,

$$\left(D\mathcal{U}^T D\mathcal{U}\right)(X) = \mathcal{P}^T(X)\begin{pmatrix} \nu_1^2(X) & 0 \\ 0 & \nu_2^2(X)\end{pmatrix}\mathcal{P}(X)$$

for some $\mathcal{P}(X) \in SO(2)$. Thus

$$\begin{aligned}\begin{pmatrix} A & B \\ B & -A\end{pmatrix} &= D\mathcal{U}^T D\mathcal{U} - \frac{1}{2}|D\mathcal{U}|^2 I \\ &= \mathcal{P}^T\begin{pmatrix}\nu_1^2 & 0 \\ 0 & \nu_2^2\end{pmatrix}\mathcal{P} - \frac{1}{2}\begin{pmatrix}\nu_1^2+\nu_2^2 & 0 \\ 0 & \nu_1^2+\nu_2^2\end{pmatrix},\end{aligned}$$

and so

$$\begin{pmatrix}\nu_1^2-\nu_2^2 & 0 \\ 0 & \nu_2^2-\nu_1^2\end{pmatrix} = 2\,\mathcal{P}\begin{pmatrix} A & B \\ B & -A\end{pmatrix}\mathcal{P}^T \in L^p_{\text{loc}}(\Omega;\mathcal{M}^{2\times 2}), \tag{2.22}$$

with the same estimate for $\nu_1^2-\nu_2^2$ as those given by Lemma 2.6 for A and B. Since $\nu_1, \nu_2 \ge 0$ and in light of (2.18),

$$|\nu_1-\nu_2| + \sqrt{\frac{|D\mathcal{U}|^2}{2}-|d|} \le 2\sqrt{|\nu_1^2-\nu_2^2|}, \tag{2.23}$$

and estimates (2.20) and (2.21) follow. □

3. Proof of the theorem. Assume (2.4) without loss of generality. Adding $M\Delta u$ to both sides of (2.1.1) and $M\Delta v$ to both sides of (2.1.2) gives the divergence-form equations

$$\begin{aligned}\Delta u &= \frac{1}{M+1}\bigtriangledown\cdot\langle Mu_x - h'(d)v_y,\, Mu_y + h'(d)v_x\rangle \equiv \bigtriangledown\cdot F, \\ \Delta v &= \frac{1}{M+1}\bigtriangledown\cdot\langle Mv_x + h'(d)u_y,\, Mv_y - h'(d)u_x\rangle \equiv \bigtriangledown\cdot G.\end{aligned} \tag{3.1}$$

The result will follow by the L^p-boundedness of Riesz potentials if appropriate L^p-estimates for the components of F and G can be obtained for all $1 < p < \infty$ (cf. [8] and [4]). The proof utilizes a bootstrapping argument. The first step is proved for $|F|$ in detail, and $|G|$ follows similarly. The bootstrap for the next step is then outlined to complete the proof.

Let $\Omega' \subset\subset \Omega$. From the integrability of $f(d)$ given in (2.9), Lemma 2.7 gives

$$\|\nu_1 - \nu_2\|_{L^{\frac{2}{1-\beta}}(\Omega')} \leq C\left(1 + \|f(d)\|_{L^{\frac{1}{1-\beta}}(\Omega)} + \|D\mathcal{U}\|_{L^2(\Omega)}\right), \tag{3.2}$$

$$\left\|\sqrt{\frac{|D\mathcal{U}|^2}{2} - |d|}\right\|_{L^{\frac{2}{1-\beta}}(\Omega')} \leq C\left(1 + \|f(d)\|_{L^{\frac{1}{1-\beta}}(\Omega)} + \|D\mathcal{U}\|_{L^2(\Omega)}\right). \tag{3.3}$$

Set

$$\begin{aligned} S_0 &= \{X \in \Omega' : |d(X)| < 1\}, \\ S_+ &= \{X \in \Omega' : d(X) \geq 1\}, \\ S_- &= \{X \in \Omega' : d(X) \leq -1\}. \end{aligned} \tag{3.4}$$

By means of singular decomposition, there exist $\mathcal{P}(X), \mathcal{Q}(X) \in SO(2)$ such that

$$D\mathcal{U} = \mathcal{P}\begin{pmatrix} \nu_1 & 0 \\ 0 & \nu_2 \end{pmatrix}\mathcal{Q} \quad \text{on } S_+, \tag{3.5}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} D\mathcal{U} = \mathcal{P}\begin{pmatrix} \nu_1 & 0 \\ 0 & \nu_2 \end{pmatrix}\mathcal{Q} \quad \text{on } S_-. \tag{3.6}$$

Claim 1. *$Mu_x - h'(d)v_y \in L^{\frac{2}{1-\beta}}(S_+)$, with the estimate*

$$\|Mu_x - h'(d)v_y\|_{L^{\frac{2}{1-\beta}}(S_+)} \leq C\left(1 + \|D\mathcal{U}\|_{L^2(\Omega)}\right). \tag{3.7}$$

Proof. To prove this, first estimate

$$\begin{aligned} |Mu_x - h'(d)v_y| &= \frac{1}{2}\big|(M + h'(d))(u_x - v_y) + (M - h'(d))(u_x + v_y)\big| \\ &\leq \frac{1}{2}\Big[\big|(M + h'(d))\,(u_x - v_y)\big| + \big|(M - h'(d))\,(u_x + v_y)\big|\Big] \\ &\equiv \frac{1}{2}\big[(\mathrm{I}) + (\mathrm{II})\big]. \end{aligned} \tag{3.8}$$

There exist $\mathcal{P}, \mathcal{Q} \in SO(2)$, rotations by angles θ and ϕ, respectively, such that (3.5) holds. Thus

$$\begin{aligned}
u_x - v_y &= tr\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} D\mathcal{U}\right]\\
&= tr\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right]\\
&= tr\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\begin{pmatrix}\cos\phi & \sin\phi\\ -\sin\phi & \cos\phi\end{pmatrix}\right] \qquad (3.9)\\
&= tr\left[\begin{pmatrix}\cos\theta & \sin\theta\\ \sin\theta & -\cos\theta\end{pmatrix}\begin{pmatrix}\nu_1\cos\phi & \nu_1\sin\phi\\ -\nu_2\sin\phi & \nu_2\cos\phi\end{pmatrix}\right]\\
&= (\nu_1-\nu_2)(\cos\theta\cos\phi+\sin\theta\sin\phi)\\
&= (\nu_1-\nu_2)\cos(\theta-\phi).
\end{aligned}$$

Still confined to S_+, since $d \geq 1$, estimate

(3.10)

$$\begin{aligned}
&(\mathrm{II}) \leq (M - h'(d))\sqrt{2}\sqrt{|D\mathcal{U}|^2} \leq K|d|^{-\beta}\left(\sqrt{2}\sqrt{|D\mathcal{U}|^2 - 2|d| + 2|d|}\right)\\
&\quad\leq 2K|d|^{-\beta}\sqrt{\frac{|D\mathcal{U}|^2}{2} - |d|} + 2K|d|^{-\beta}\sqrt{|d|} \leq 2K\sqrt{\frac{|D\mathcal{U}|^2}{2} - |d|} + 2K|d|^{\frac{1-\beta}{2}}.
\end{aligned}$$

From estimates (2.9), (2.20), and (2.21), we get

$$\begin{aligned}
\|\mathrm{I}\|_{L^{\frac{2}{1-\beta}}(S_+)} + \|\mathrm{II}\|_{L^{\frac{2}{1-\beta}}(S_+)} &\leq C\left(1 + \|f(d)\|_{L^{\frac{1}{1-\beta}}(\Omega)} + \|D\mathcal{U}\|_{L^2(\Omega)}\right)\\
&\leq C\left(1 + \|D\mathcal{U}\|_{L^2(\Omega)}\right).
\end{aligned} \qquad (3.11)$$

This completes the proof of Claim 1.

Claim 2. *$Mu_x - h'(d)v_y \in L^{\frac{2}{1-\beta}}(S_-)$, with the estimate*

$$\|Mu_x - h'(d)v_y\|_{L^{\frac{2}{1-\beta}}(S_-)} \leq C\left(1 + \|D\mathcal{U}\|_{L^2(\Omega)}\right). \qquad (3.12)$$

Proof. The proof is nearly the same as that of Claim 1, except that the respective methods for estimating (I) and (II) are reversed. In S_-, estimate

(3.13)

$$\begin{aligned}
&(\mathrm{I}) \leq (M + h'(d))\sqrt{2}\sqrt{|D\mathcal{U}|^2} \leq K|d|^{-\beta}\left(\sqrt{2}\sqrt{|D\mathcal{U}|^2 - 2|d| + 2|d|}\right)\\
&\quad\leq 2K|d|^{-\beta}\sqrt{\frac{|D\mathcal{U}|^2}{2} - |d|} + 2K|d|^{-\beta}\sqrt{|d|} \leq 2K\sqrt{\frac{|D\mathcal{U}|^2}{2} - |d|} + 2K|d|^{\frac{1-\beta}{2}}.
\end{aligned}$$

As for (II), using (3.6), compute as before

(3.14)

$$\begin{aligned}
u_x + v_y = tr\, D\mathcal{U} &= tr\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}^2 D\mathcal{U}\right] = tr\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right]\\
&= (\nu_1-\nu_2)\cos(\theta-\phi).
\end{aligned}$$

Thus

$$\|\mathrm{I}\|_{L^{\frac{2}{1-\beta}}(S_-)} + \|\mathrm{II}\|_{L^{\frac{2}{1-\beta}}(S_-)} \leq C\left(1 + \|D\mathcal{U}\|_{L^2(\Omega)}\right). \qquad (3.15)$$

Claim 3. *$Mu_y + h'(d)v_x \in L^{\frac{2}{1-\beta}}(S_+)$, along with an estimate of the form* (3.7) *for $Mu_y + h'(d)v_x$.*

Proof. As in the proofs of Claims 1 and 2, first estimate this quantity as follows:

$$\begin{aligned}|Mu_y + h'(d)v_x| &\leq \frac{1}{2}\Big[|(M + h'(d))(u_y + v_x)| + |(M - h'(d))(u_y - v_x)|\Big] \\ &\equiv \frac{1}{2}\,[(\mathrm{III}) + (\mathrm{IV})].\end{aligned} \tag{3.16}$$

Since $d \geq 1$, it follows that

$$u_y + v_x = (D\mathcal{U})_{12} + (D\mathcal{U})_{21} = \left[\left(\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right)_{12} + \left(\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right)_{21}\right].$$

Reading off these entries from the matrix

$$\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q} = \begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\begin{pmatrix}\cos\phi & \sin\phi\\ -\sin\phi & \cos\phi\end{pmatrix}$$

gives

$$\begin{aligned}u_y + v_x &= (\nu_1\cos\theta\sin\phi + \nu_2\sin\theta\cos\phi - \nu_1\sin\theta\cos\phi - \nu_2\cos\theta\sin\phi) \\ &= (\nu_1 - \nu_2)(\cos\theta\sin\phi - \sin\theta\cos\phi) \\ &= (\nu_1 - \nu_2)\sin(\phi - \theta).\end{aligned} \tag{3.17}$$

The estimate for (IV) is the same as for (II) in Claim 1, and so Claim 3 is proved.

Claim 4. *$Mu_y + h'(d)v_x \in L^{\frac{2}{1-\beta}}(S_-)$, along with an estimate of form* (3.14) *for $Mu_y + h'(d)v_x$.*

Proof. Again, the proof is nearly the same as that of Claim 3, except for the reversal of the roles of (III) and (IV). The estimate for (I) from Claim 2 holds for (III) here. To estimate (IV), write

$$\begin{aligned}u_y - v_x &= \left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}D\mathcal{U}\right]_{12} + \left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}D\mathcal{U}\right]_{21} \\ &= \left\{\left[\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right]_{12} + \left[\mathcal{P}\begin{pmatrix}\nu_1 & 0\\ 0 & \nu_2\end{pmatrix}\mathcal{Q}\right]_{21}\right\} \\ &= (\nu_1 - \nu_2)\sin(\phi - \theta).\end{aligned} \tag{3.18}$$

This completes the proof of Claim 4.

Claim 5. *$D\mathcal{U} \in L^{\frac{2}{1-\beta}}(S_0)$, with an estimate*

$$\|D\mathcal{U}\|_{L^{\frac{2}{1-\beta}}(S_0)} \leq C\left(1 + \|D\mathcal{U}\|_{L^2(\Omega)}\right). \tag{3.19}$$

Proof. The estimate follows from (2.21) and the observation that the estimate $|D\mathcal{U}| \leq \sqrt{|D\mathcal{U}|^2 - 2|d|} + \sqrt{2}$ holds on S_0, and Claim 5 is proved.

We have shown so far that $\Delta u = \bigtriangledown \cdot F$, where $|F| \in L^p(\Omega')$ for $p = 2/(1-\beta)$. An argument similar to the one given above shows that $|G| \in L^p(\Omega')$ as well. Then from local estimates for Riesz potentials comes the estimate

$$\|D\mathcal{U}\|_{L^{\frac{2}{1-\beta}}(\Omega'')} \leq C\left(1 + \|\mathcal{U}\|_{L^2(\Omega)} + \|D\mathcal{U}\|_{L^2(\Omega)}\right) \tag{3.20}$$

on any subdomain $\Omega'' \subset\subset \Omega'$. Now the proof employs the bootsrapping. From (2.8), it is easy to see that

$$|f(d)| \leq C(1+|D\mathcal{U}|)^{2(1-\beta)}. \tag{3.21}$$

From (3.20), we then get

$$\|f(d)\|_{L^{\frac{1}{(1-\beta)^2}}(\Omega'')} \leq C\left(1+\|\mathcal{U}\|_{L^2(\Omega)}+\|D\mathcal{U}\|_{L^2(\Omega)}\right). \tag{3.22}$$

After passing to further subdomains, Lemmas 2.6 and 2.7 apply with $p=(1-\beta)^2$, and estimates (3.7), (3.12), (3.19), and (3.20) follow with $(1-\beta)^2$ replacing $(1-\beta)$ and with the modification that the $\|\mathcal{U}\|_{L^2(\Omega)}$ term be included in the right-hand sides of all these estimates. After m steps, $(1-\beta)$ is replaced by $(1-\beta)^m$, and the theorem follows. □

Remark. In the case where $\beta > 1$, $f(d) \in L^\infty(\Omega)$ and no bootstrap is required.

REFERENCES

[1] J. M. Ball and F. Murat, $W^{1,p}$ *Quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal., 58 (1984), pp. 225–253.

[2] P. Bauman, N. C. Owen, and D. Phillips, *Maximum principles and a priori estimates for a class of problems from nonlinear elasticity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 119–157.

[3] M. Chipot and L. Evans, *Linearisation at infinity and Lipschitz estimates for certain problems in the calculus of variations*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 291–303.

[4] E. DiBenedetto and J. Manfredi, *On the higher integrability of the gradient of weak solutions of certain degenerate elliptic systems*, Amer. J. Math., 115 (1993), pp. 1107–1134.

[5] M. Dougherty and D. Phillips, *Higher integrability of equilibria for certain rank-one convex integrals*, SIAM J. Math. Anal., 28 (1997), pp. 270–273.

[6] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.

[7] M. Giaquinta and E. Giusti, *On the regularity of the minima of variational integrals*, Acta Math., 148 (1982), pp. 31–46.

[8] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. Math. Anal.
Vol. 28, No. 3, pp. 539–569, May 1997

STABILITY OF GASEOUS STARS IN SPHERICALLY SYMMETRIC MOTIONS[*]

SONG-SUN LIN[†]

Abstract. We study the linearized stability of stationary solutions of gaseous stars which are in spherically symmetric and isentropic motion. If viscosity is ignored, we have following three types of problems: (EC), Euler equation with a solid core; (EP), Euler–Poisson equation without a solid core; (EPC), Euler–Poisson equation with a solid core. In Lagrangian formulation, we prove that any solution of (EC) is neutrally stable. Any solution of (EP) and (EPC) is also neutrally stable when the adiabatic index $\gamma \in (\frac{4}{3}, 2)$ and unstable for (EP) when $\gamma \in (1, \frac{4}{3})$. Moreover, for (EPC) and $\gamma \in (1, 2)$, any solution with small total mass is also neutrally stable. When viscosity is present ($\nu > 0$), the velocity disturbance on the outer surface of gas is important. For $\nu > 0$, we prove that the neutrally stable solution (when $\nu = 0$) is now stable with respect to positive-type disturbances, which include Dirichlet and Neumann boundary conditions. The solution can be unstable with respect to disturbances of some other types. The problems were studied through spectral analysis of the linearized operators with singularities at the endpoints of intervals.

Key words. stability, isentropic gas, self-gravitating, solid core, limit-point singularity

AMS subject classifications. 35J65, 35P30, 85A15, 85A20

PII. S0036141095292883

1. Introduction.

In this paper, we shall study the stability problem of gaseous stars which are in spherically symmetric and isentropic motion. The problem originated in Newtonian (nonrelativistic) astrophysical theory. A model equation for describing such motion is shown below:

$$\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial r} + \rho\frac{\partial v}{\partial r} + \frac{2}{r}\rho v = 0, \tag{1.1}$$

$$\begin{aligned}\rho\left(\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial r}\right) + \frac{\partial p}{\partial r} = &- \frac{\rho}{r^2}\left\{M_0 + 4\pi\delta\int_{R_0}^{r}\rho(t,s)s^2ds\right\}\\ &+ \nu\left\{\frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} - \frac{2}{r^2}v\right\},\end{aligned} \tag{1.2}$$

$$p = A\rho^{\gamma}, \tag{1.3}$$

where $t \geq 0$ and $0 \leq R_0 < r < \infty$; see, e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14, 20]. Here the unknown variable ρ is the density of the gas and v is the outward velocity. p is the pressure, A is a positive constant which is related to entropy, and $\gamma \in$ (1,2) is the adiabatic exponent.

The explanation of the physical parameters δ, M_0, R_0, and ν is as follows:

δ is the effect of self-gravitating of gas, the mutual graviational attraction among gas molecules, and is assumed to be either 0 or 1. If $\delta = 0$, we ignore the effect of self-gravitating. This may happen when the total amount of gas is relatively small. If $\delta = 1$, we then consider the self-gravitating of gas to be important.

[*] Received by the editors October 6, 1995; accepted for publication January 31, 1996. This research was partially supported by the National Science Council of the Republic of China.
http://www.siam.org/journals/sima/28-3/29288.html

[†] Department of Applied Mathematics, National Chiao Tung University, Hsin-chu 30050, Taiwan, Republic of China (sslin@math.nctu.edu.tw).

M_0 is the total mass of the solid core surrounded by the gas. If $M_0 = 0$, then we assume that $R_0 = 0$. This is the case when there is no solid core and also no vacuum in the central part of the gaseous body. If $M_0 > 0$, we assume that there is a stationary, spherical solid core surrounded by the gas. In this case, we normalize the radius of the solid core with $R_0 = 1$. We also assume that the gas is in contact the surface of the solid core, i.e., no vacuum exists between the core and the gas. A nonslip condition is now imposed at the interface, i.e.,

$$v(t,1) = 0 \quad \text{for } t \geq 0. \tag{1.4}$$

We note that astrophysicists consider the solid core to be made of condensed gases in which there may be complicated activity that influences the surrounding gas. However, for mathematical simplicity, we will consider these condensed gases to be a solid core and ignore their influence on the surface gas.

ν is viscosity coefficient. We are mainly concered with inviscid flow, i.e., $\nu = 0$. After presenting a detailed study of inviscid flow, we will discuss the effect of viscosity on the stability of stationary solutions.

If viscosity is ignored, then according to the different combinations of parameters δ, M_0, and R_0, we have following three types of problems:

(EC): Euler equation with solid core ($\delta = 0,\ M_0 > 0,\ R_0 = 1,\ \nu = 0$);

(EP): Euler–Poisson equation without solid core ($\delta = 1$, $M_0 = 0$, $R_0 = 0$, $\nu = 0$);

(EPC): Euler–Poisson equation with solid core ($\delta = 1$, $M_0 > 0$, $R_0 = 1$, $\nu = 0$).

If viscosity is present, i.e., $\nu > 0$, then the Euler equation will be replaced by a Navier–Stokes equation and we have problems (NSC), (NSP), and (NSPC), respectively.

The stationary solution $(\overline{\rho}(r), 0)$ of (1.1)–(1.3) satisfies

$$\frac{d\overline{p}}{dr} = -\frac{\overline{\rho}}{r^2}\left\{M_0 + 4\pi\delta\int_{R_0}^{r}\overline{\rho}(s)s^2\right\}. \tag{1.5}$$

If we introduce the variable $u(r)$ and the parameter $\mu > 0$ in

$$\overline{\rho} = C_\gamma u^q \quad \text{and} \quad \mu = d_\gamma M_0,$$

where

$$q = \frac{1}{\gamma - 1}, \qquad C_\gamma = \left\{\frac{A\gamma}{4\pi(\gamma-1)}\right\}^{\frac{1}{2-\gamma}}, \quad \text{and} \quad d_\gamma = \left\{(4\pi)^{\gamma-1}\frac{\gamma-1}{A\gamma}\right\}^{\frac{1}{2-\gamma}},$$

then (1.5) and (1.4) can be studied by considering the following initial-value problems: for (EC),

$$\left.\begin{array}{l} u'' + \frac{2}{r}u' = 0, \quad r > 1, \\ u(1,\alpha,\mu) = \alpha \quad \text{and} \quad u'(1,\alpha,\mu) = -\mu \end{array}\right\}; \tag{1.6}$$

for (EP),

$$\left.\begin{array}{l} u'' + \frac{2}{r}u' + u^q = 0, \quad r > 0, \\ u(0,\alpha) = \alpha \quad \text{and} \quad u'(0,\alpha) = 0 \end{array}\right\}; \tag{1.7}$$

and for (EPC),

$$\left.\begin{array}{l} u'' + \frac{2}{r}u' + u^q = 0, \quad r > 1, \\ u(1,\alpha,\mu) = \alpha \quad \text{and} \quad u'(1,\alpha,\mu) = -\mu \end{array}\right\}. \tag{1.8}$$

Here $\alpha > 0$ is taken as a shooting parameter.

The total mass of the stationary solution u is given by

$$\tilde{M}(u) = 4\pi C_\gamma \int_{R_0}^{R} u^q(r) r^2 dr, \tag{1.9}$$

where $R \in (R_0, \infty]$ is the first zero of u, i.e.,

$$u(R) = 0 \quad \text{and} \quad u(r) > 0 \quad \text{in } (R_0, R).$$

From a physical point of view, we are only interested in a stationary solution with finite total mass.

The solution of (1.6) with finite total mass can be written explicitly as

$$u = \mu \left(\frac{1}{r} - \frac{1}{R} \right) \tag{1.10}$$

for some $R \in (1, \infty]$.

The solution of (1.7) has been studied extensively by Lane et al.; see, e.g., [1]. Their solutions include the ball type ($R < \infty$), the ground-state type ($R = +\infty$), and the singularity type, i.e., $\lim_{r \to 0^+} u(r) = \infty$.

Equation (1.8) has recently been studied in [5] and may have multiple solutions for certain μ and $\tilde{M}$ when $q > 3$.

The multiplicity results of these problems will be given in section 2.

In this paper, we mainly study the stability of stationary solutions obtained from (1.6), (1.7), and (1.8) since only the local existence and not the global-existance of the initial-value problem in (1.1)–(1.3) is known (see, e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14]). We therefore need only study the linearized stability of these stationary solutions.

The linearized stability problem of the stationary solution $\overline{\rho}(r)$ will be studied in Lagrangian formulation. Indeed, equations (1.1)–(1.3) can be written in Lagrangian coordinates as

$$\rho_t + 4\pi\rho(r^2 v)_x = 0, \tag{1.11}$$

$$v_t + 4\pi r^2 p_x + \frac{1}{r^2}(M_0 + x) = 16\pi^2 \nu (r^2 \rho v_x)_x - 2\nu v (r^2 \rho)^{-1}, \tag{1.12}$$

$$r = \left\{ R_0 + \frac{3}{4\pi} \int_0^x \frac{1}{\rho(t, y)} dy \right\}^{\frac{1}{3}} \quad \text{and} \quad x = 4\pi \int_{R_0}^{r} \rho(s, t) s^2 ds,$$

where $t \geq 0$ and $x \in (0, \tilde{M})$. We assume that the perturbation of $(\overline{\rho}(x), 0)$ is in a radial direction only and write

$$\rho(t, x) = \overline{\rho}(x)\{1 + \varepsilon e^{\lambda t} \Phi(x)\} \quad \text{and} \quad v(t, x) = \varepsilon e^{\lambda t} \Psi(x) \tag{1.13}$$

in (1.11) and (1.12), where $|\varepsilon|$ is small. Let

$$\phi(x) = \int_0^x \frac{\Phi(y)}{\overline{\rho}(y)} dy.$$

Then the linear equations for Φ and Ψ can be simplified as follows:

$$(\overline{\rho}\overline{p}\phi_x)_x - \frac{1}{\pi\gamma\overline{r}^3}\overline{p}_x\phi = \frac{\lambda^2}{\gamma(4\pi\overline{r}^2)^2}\phi - \frac{\lambda\nu}{4\pi\gamma\overline{r}^2}\left\{16\pi^2\left[\overline{r}^4\overline{\rho}\left(\frac{1}{4\pi\overline{r}^2}\phi\right)_x\right]_x - \frac{2}{\overline{r}^2\overline{\rho}}\frac{1}{4\pi\overline{r}^2}\phi\right\} \tag{1.14}$$

with boundary condition

$$\phi(0) = 0, \tag{1.15}$$

where

$$\overline{r} = \left\{1 + \frac{3}{4\pi}\int_0^x \frac{1}{\overline{\rho}(y)}dy\right\}.$$

Transforming (1.14) into $\overline{r}$-coordinates and writting $\phi(x) = \psi(\overline{r})$, we obtain

$$\overline{L}\psi \equiv (\overline{r}^{-2}\overline{p}\psi')' - \frac{4}{\gamma}\overline{r}^{-3}\overline{p}'\psi = \frac{\lambda^2}{\gamma}\overline{r}^{-2}\overline{\rho}\psi - \frac{\lambda\nu}{\gamma}(\overline{r}^{-2}\psi')' \tag{1.16}$$

with $\psi(R_0) = 0$, where $\overline{p}$ is the pressure in $\overline{r}$-coordinates. Since $\overline{\rho}(R) = 0$, if $\nu = 0$, then (1.16) is singular at $\overline{r} = R$. We can prove that the singularity at R is a limit-point type and so $\overline{L}$ is self-adjoint. Therefore, λ^2 is real for any eigenvalue λ when $\nu = 0$. Now $\overline{\rho}$ is called neutrally stable if $\lambda^2 < 0$ for any eigenvalue λ and unstable if $\lambda_1^2 > 0$ for some eigenvalue λ_1. Hence if $\nu = 0$, then neutrally stable is the best we can hope for. Indeed, when $\nu = 0$, we have our stability results for ball-type solutions as follows.

THEOREM 1.1. *Assume that $\nu = 0$ and ball-type solutions have been considered. Then*

(I) *any solution of* (EC) *is neutrally stable*;

(II) *any solution of* (EP) *is neutrally stable if $q \in (1, 3)$ and unstable if $q > 3$; and*

(III) *for* (EPC), *we have the following:*

(i) *any solution is neutrally stable if $q \in (1, 3]$,*

(ii) *for any $q > 1$, $u(\cdot, \alpha, \mu)$ is neutrally stable if $\alpha \in (0, \mu]$, and*

(iii) *if $|R - 1|$ is sufficiently small, then it is neutrally stable.*

Some stability results concerning ground-state-type and singularity-type solutions are also presented in section 4.

When viscosity is present and $\lambda \notin [-\frac{\gamma}{\nu}p(R_0), 0]$, then (1.16) is regular at R. In this case, the viscosity term plays the dominant role in studying the eigenvalue problems. Now $\overline{\rho}$ is called stable if $\mathrm{Re}\lambda < 0$ for any eigenvalue λ and unstable if $\mathrm{Re}\lambda_1 > 0$ for some eigenvalue λ_1. Note that (1.16) is genuinely quadratic in λ (linear in λ^2 when $\nu = 0$) and λ is complex in general. Hence when $\nu > 0$, we may have better than the neutral stability that we have when $\nu = 0$. Since the outer surface of gas is a free surface, the velocity disturbance Ψ on it will play an important role. For example, we have stability results for (EC), (EP), and (EPC) as follows.

THEOREM 1.2. *Let u be a neutrally stable, ball-type stationary solution of* (EC), (EP), *or* (EPC) *when $\nu = 0$. Then for any $\nu > 0$, u is stable with respect to $\Psi = \psi_1 + i\psi_2$ if $\psi_j(\tilde{M})\psi_j'(\tilde{M}) \leq 0$ for both $j = 1$ and 2 on the gas surface. On the other hand, there is a positive constant κ_* depending on u such that u is unstable with respect to $\Psi = \psi_1 + i\psi_2$ if $\psi_1'(\tilde{M})/\psi_1(\tilde{M}) \geq \kappa_*$ and some ψ_2.*

The precise definition of stability with respect to the boundary disturbance Ψ is given in section 5.

The paper is organized as follows. In section 2, we recall some useful multiplicity results for stationary solutions with finite total masses. Their stabilities will be investigated in subsequent sections. In section 3, we study the linearized operators $\overline{L}$ and prove that they have limit-point-type singularities at their endpoints. We also provide a useful comparison lemma to test for stability. In section 4, we prove various stability results, which include Theorem 1.1. The solutions for other types of stability problems are also studied. In section 5, we study the effect of viscosity on stability problems and prove some results, including Theorem 1.2. In Appendix A, we study the asymptotic behavior of solutions of (1.16) at R when $\nu = 0$, which is very useful for studying ball-type solutions. In Appendix B, we recall Friedrichs' criteria for the spectrum discreteness of differential operators that have singular endpoints. These criteria are very useful in studying ground-state-type and singularity-type solutions.

2. Stationary solutions. In this section, we recall some multiplicity results for stationary solutions without interior vacuums and with finite total masses. Let $R \leq \infty$ be the first zero of solution u and $\tilde{M}(u)$ be the total mass given in (1.9). For notational simplicity, we omit the constant $4\pi C_\gamma$ in (1.9) and then define

$$M(u) = \int_{R_0}^{R} u(r)^q r^2 dr, \tag{2.1}$$

where $R_0 = 0$ for (EP) and $R_0 = 1$ for (EC) and (EPC).

Since the total mass of a gas remains constant while it is in motion and it may tend to a stationary state as time goes by, it is useful to know the numbers of stationary solutions for the same total mass. Hence we try to answer the following questions.

Questions. Given $M > 0$, how many solutions u are there for (EP) with $M(u) = M$? Given $\mu > 0$ and $M > 0$, how many solutions u are there for (EC) or (EPC) with $M(u) = M$?

Complete answers of (EC) and (EP) can be provided; see, e.g., [1]. However, (EPC) has only recently been studied and the result is complete for $1 < q \leq 3$ but partial when $q > 3$; see [5].

First, for (EC), the solution of (1.6) is given by

$$u(r, \alpha, \mu) = \alpha - \mu + \mu \frac{1}{r}. \tag{2.2}$$

If $\alpha \in (0, \mu)$, then $u(R(\alpha, \mu), \alpha, \mu) = 0$ with

$$R(\alpha, \mu) = \left(1 - \frac{\alpha}{\mu}\right)^{-1}.$$

In this case, we may write $u(\cdot, \alpha, \mu) = u_{R,\mu}$ with

$$u_{R,\mu}(r) = \mu \left(\frac{1}{r} - \frac{1}{R}\right).$$

It is clear that $M(u_{R,\mu})$ is strictly increasing in R and tends to

$$M_q^* = \begin{cases} +\infty & \text{if } 1 < q \leq 3, \\ \dfrac{1}{q-3} \cdot \mu^q & \text{if } q > 3. \end{cases}$$

If $\alpha = \mu$, then $R(\mu, \mu) = +\infty$ and

$$u(r, \mu, \mu) = \frac{\mu}{r}$$

with

$$M(u(\cdot, \mu, \mu)) = M_q^*.$$

If $\alpha > \mu$, then $M(u(\cdot, \alpha, \mu)) = \infty$, which is not of physical interest. Hence we have the following unique result for (EC).

PROPOSITION 2.1. *For any $q > 1$, $\mu > 0$, and $M \in (0, M_q^*)$, there is a unique solution $u_{R,\mu}$ for* (EC) *such that $M(u_{R,\mu}) = M$.*

Next, for (EP), we consider the initial-value problem

$$u'' + \frac{2}{r}u' + u^q = 0, \quad r > 0, \tag{2.3}$$

$$u'(0, \alpha) = 0 \quad \text{and} \quad u(0, \alpha) = \alpha > 0. \tag{2.4}$$

It is known that solutions of (2.3) have similar properties. Indeed, if $u(r)$ is a solution of (2.3), then for any $\beta > 0$,

$$u_\beta(r) = \beta^\sigma u(\beta r) \tag{2.5}$$

is also a solution, where $\sigma = \frac{2}{q-1}$. The total mass of u_β is

$$M(u_\beta) = \beta^{\frac{3-q}{q-1}} M(u). \tag{2.6}$$

The property (2.5) is related to the following classical Lane–Emden–Fowler transformations:

Let

$$r = e^{-\tau} \quad \text{and} \quad z(\tau) = r^\sigma u(r). \tag{2.7}$$

(2.3) can then be transformed into the autonomous equation

$$z'' + (2\sigma - 1)z' + \sigma(\sigma - 1)z + z^q = 0 \tag{2.8}$$

or, equivalently, the dynamic system

$$\begin{cases} z' = y, \\ y' = -\{2\sigma - 1)y + \sigma(\sigma - 1)z + z^q\}. \end{cases} \tag{2.9}$$

If $q \in (1, 3]$, then $0 = (0, 0)$ is the only equilibrium for (2.9) on the right half-plane $R_+^2 = \{(z, y) : z \geq 0\}$. If $q > 3$, then there is another equilibrium $S = (z_\sigma, 0)$, where

$$z_\sigma = \{\sigma(1 - \sigma)\}^{\frac{2}{\sigma}}. \tag{2.10}$$

0 is always a saddle point with the unstable manifold Γ, which is leaving in the direction $(1, 1 - \sigma)^t$, and the stable manifold $\tilde{\Gamma}$, which is arriving for the direction $(1, -\sigma)$, where $(a, b)^t$ is the transpose of vector (a, b) in R^2. Let

$$q^+ = 1 + \frac{2}{\sigma^+} \quad \text{and} \quad \sigma^+ = \sqrt{2} - \frac{1}{2}.$$

It is than easy to verify that $q^+ \in (3,5)$.

We now list some useful properties of the equilibrium S and system (2.9) on the phase plane R^2_+.

Proposition 2.2.

(I)

(i) *If* $q \in (3, q^+)$, *then* S *is a stable improper node.*

(ii) *If* $q = q^+$, *then* S *is a stable proper node.*

(iii) *If* $q \in (q^+, 5)$, *then* S *is stable spiral.*

(iv) *If* $q = 5$, *then* S *is a center.*

(v) *If* $q > 5$, *then* S *is an unstable spiral.*

(II)

(i) *For* $q \in (3,5)$, *the unstable manifold* Γ *of* 0 *is a heteroclinic orbit connecting* 0 *and* S. *There is no nontrivial periodic orbit on* R^2_+.

(ii) *For* $q = 5, \Gamma = \tilde{\Gamma}$, *i.e.,* Γ *is a homoclinic orbit of* 0. *The inside of* Γ *is covered by a family of concentric periodic orbits centered around* 0.

(iii) *For* $q > 5$, *the stable manifold* $\tilde{\Gamma}$ *of* 0 *is a heteroclinic orbit connecting* 0 *and* S.

The proofs are elementary and omitted; see [1] for details.

Every trajectory in the phase plane of (2.9) represents a family of self-similar solutions in (2.5). After carefully investigating the trajectories in the phase plane, we have exactly four types of solutions for (EP) with finite total mass for (EP):

(i) B-type solutions: ball-type solutions that lie on $\tilde{\Gamma}$ and appear when $q \in (1,5)$;

(ii) G-type solutions: ground-state solutions that also lie on $\tilde{\Gamma}$ and only appear when $q \geq 5$; they also have fast decay rates as $r \to +\infty$, i.e.,

$$\lim_{r \to +\infty} ru(r) \in (0,\infty); \tag{2.11}$$

(iii) SB-type solutions: ball-type solutions with a singularity at $r = 0$ that appear when $q \in (3,5)$ and are trajectories between Γ and $\tilde{\Gamma}$ that have a weak singularity, i.e., u satisfies

$$\lim_{r \to 0^+} r^\sigma u(r) \in (0,\infty); \tag{2.12}$$

(iv) SG-type solutions: ground-state solutions with a singularity at $r = 0$ that lie on Γ and satisfy (2.12); they also appear when $q \in (3,5)$;

Note that if the singularity at $r = 0$ is strong, i.e.,

$$\lim_{r \to 0^+} ru(r) > 0, \tag{2.13}$$

then u has an infinite total mass: for example, the SB-type solution when $q \in (1,3)$. If the ground-state solution has a slow-decay rate at ∞, i.e.,

$$\lim_{r \to \infty} r^\sigma u(r) > 0, \tag{2.14}$$

then u also has an infinite total mass, which includes the following cases:

(i) $q_\sigma r^{-\sigma}$ for $q > 3$; this corresponds to the equilibrium $S = (z_\sigma, 0)$;

(ii) when $q = 5$, all trajectories lie on homoclinic orbit Γ;

(iii) when $q > 5$, all trajectories spiral out from S;

With this preparation complete, we can now state our unique results for (EP).

Proposition 2.3. *For* (EP), *we have the following:*

(i) *If $q \in (1,3)$ and any $M > 0$, there is a unique—B-type—solution u such that $M(u) = M$.*

(ii) *If $q = 3$, only a special $\hat{M}$ of a stationary solution u—a B-type solution—admits.* (*All similar solutions of u also have the same total mass $\hat{M}$.*)

(iii) *If $q \in (3,5)$ and any $M > 0$, there are unique B-type, SB-type, and SG-type solutions with the same total mass M.*

(iv) *If $q = 5$ and any $M > 0$, there is a unique—G-type—solution u such that $M(u) = M$.*

(v) *If $q > 5$, there is no stationary solution with finite total mass.*

Proof. The proofs are based on the phase-plane analysis in (2.9) and the use of (2.6), and they are elementary. Thus the details are omitted. □

As for (EPC), there are two types of solutions with finite total mass:

(i) BC-type solutions: ball-type solutions with solid cores;

(ii) GC-type solutions: ground-state solutions with solid cores that satisfy (2.11).

We recall some results from [5].

Proposition 2.4. *For* (EPC), *we have the following:*

(i) *When $q \in (1,3]$, for any $\mu > 0$ and $M > 0$, there is a unique—BC-type—solution u that satisfies $M(u) = M$.*

(ii) *When $q > 3$, for any $\mu > 0$, the solution set is the disjoint union of N many connected components $C_k = \{u(\cdot, \alpha, \mu) : \alpha \in (\tilde{\alpha}_k,\, \hat{\alpha}_k)\}$, $k = 1, 2, \ldots, N$, where $N = N(\mu, q)$ is a positive integer or infinity.*

At C_k with $k \geq 2$, $M((u(\cdot, \alpha, \mu))$ tends to infinity at at least one end. At $C_1, \tilde{\alpha}_1 = 0$ and $\hat{\alpha}_1 > \mu$.

For detailed statements of Proposition 2.4(ii), see Theorems 3.5, 3.7, 3.9, and 3.13 in [5].

Remark 2.5. When there is a vaccum in the central part of the gaseous body that is also stationary, then u satisfies

$$u''(r) + \frac{2}{r}u'(r) + u^q(r) = 0, \qquad R_1 < r < R_2, \tag{2.15}$$

$$u(R_1) = 0 = u(R_2), \tag{2.16}$$

where $0 < R_1 < R_2 \leq \infty$. For any $q > 1$ and $0 < R_1 < R_2 < \infty$, Ni and Nussbaum [17] proved that there is a unique positive solution of (2.15) and (2.16). In contrast to Proposition 2.3(v), for any $q > 1$, the solution u of (2.15) and (2.16) with $R_2 < \infty$ has a finite total mass. We can then ask the following questions: Given $q > 1$ and $M > 0$, how many solutions u are there for (2.15) and (2.16) with $M(u) = M$? What is the stability of these annular-type solutions? These problems will be studied later.

3. Linearizations.

In this section, we will use a Lagrangian formulation to study the stability of the stationary solutions obtained in last section. Since we want to know the stability result when the outer surface of the gas is also perturbed, it is convenient to work in Lagrangian coordinates. We study only the inviscid flow in this section and defer study of the the viscous flow to section 5.

For notational simplicity, we replace $\bar{r}$ with r in (1.16) with $\nu = 0$. We then obtain

$$\mathcal{L}\psi = -\ell W \psi \quad \text{in } (R_0, R), \tag{3.1}$$

where

$$\mathcal{L}\psi \equiv (r^{-2}p\psi')' - \frac{4}{\gamma}r^{-3}p'\psi, \qquad W(r) \equiv \frac{\rho(r)}{\gamma r^2}, \quad \text{and} \quad \ell = -\lambda^2.$$

ψ also satisfies the boundary condition

$$\psi(R_0) = 0. \tag{3.2}$$

In terms of u, (3.1) can also be written as

$$L_0\psi \equiv \psi'' + \left\{(1+q)\frac{u'}{u} - \frac{2}{r}\right\}\psi' - \frac{4q}{r}\frac{u'}{u}\psi = -\ell(\gamma AC_\gamma^{\gamma-1})^{-1}\frac{\psi}{u}. \tag{3.3}$$

Since $u(R) = 0$, $\mathcal{L}$ is singular at R. Furthermore, $\mathcal{L}$ is also singular at $r = 0$ for (EP).

When $R < \infty$, we first study the asymptotic behavior of solution ψ of (3.1) at R. Indeed, we have the following result. (The proof is given in Appendix A.)

LEMMA 3.1. *Let $R < \infty$. If ℓ is real and ψ is a (real) solution of (3.1) in (R_0, R), then either ψ is bounded at $r = R$ or $\psi(r) = (R-r)^{-q}\hat{\psi}(r)$ for r close to R, with $\hat{\psi}(R) \neq 0$, and $\hat{\psi}$ is continuous at R. Furthermore, in the former case, ψ is C^2 at R, and in the latter case,*

$$\psi'(r) = q(R-r)^{-q-1}\hat{\psi}(R) + o((R-r)^{-q-1}) \tag{3.4}$$

as $r \to R^-$.

Similarly, if $R_0 = 0$, then either $\psi(0) \neq 0$ or $|\psi(r)| \leq Cr^3$ for r close to 0 and some $C > 0$.

To study the singularity type at R, it is convenient to remove the weight function W from right-hand side of (3.1). Indeed, if $R_0 = 1$, let $r_0 = 1$, and if $R_0 = 0$, choose any $r_0 \in (0, R)$ and fix it. Then define

$$s = s(r) = \int_{r_0}^{r} W(\tau)d\tau = \frac{1}{\gamma}\int_{r_0}^{r} \tau^{-2}\rho(\tau)d\tau \tag{3.5}$$

and

$$S_0 = \int_{r_0}^{R_0} W(\tau)d\tau \quad \text{and} \quad S = \int_{r_0}^{R} W(\tau)d\tau.$$

It is clear that $S_0 = 0$ when $R_0 = 1$ and $S_0 = -\infty$ when $R_0 = 0$. Furthermore, $W > 0$ in (R_0, R) implies that the inverse function of $s(r)$ exists. We may denote it by

$$r = r(s)$$

for $s \in (S_0, S)$. Let

$$\chi(s) = \psi(r(s)).$$

Then (3.1) is transformed into

$$\tilde{\mathcal{L}}\chi = -\ell\chi \quad \text{in } (S_0, S), \tag{3.6}$$

where

$$\tilde{\mathcal{L}}\chi = \frac{1}{\gamma}\frac{d}{ds}\left(r^{-4}p\rho\frac{d\chi}{ds}\right) - 4\frac{p'}{r\rho}\chi.$$

Let $L^2(S_0, S)$ be the complex-valued L^2-space on (S_0, S) with the standard inner product

$$(\chi, \tilde{\chi}) \equiv \int_{S_0}^{S} \overline{\chi}\tilde{\chi} ds. \tag{3.7}$$

It is clear that

$$(\chi, \tilde{\chi}) = \int_{R_0}^{R} \overline{\psi}\tilde{\psi} W(r) dr \equiv (\psi, \tilde{\psi})_w. \tag{3.8}$$

Here $(\ ,\ )_w$ defines an inner product in space $L^2_w(R_0, R)$ by (3.8).

Now we can prove $\tilde{\mathcal{L}}$ has limit-point-type singularity at S.

LEMMA 3.2. *If $S < \infty$, then $\tilde{\mathcal{L}}$ is limit-point type at S. Furthermore, for* (EP), *$\tilde{\mathcal{L}}$ is also limit-point type at $-\infty$.*

Proof. From [2], it is known that $\tilde{\mathcal{L}}$ is the limit-point-type singularity at S if we can find a solution pair $\{\ell, \chi\}$ for (3.6) in a neighborhood of S such that χ is not L^2. This can be done as follows:

Since

$$p = AC_\gamma^\gamma u^{q+1} \tag{3.9}$$

and

$$p' = A(q+1)C_\gamma^\gamma u^q u',$$

we have

$$\frac{p'(r)}{r\rho(r)} = A(q+1)C_\gamma^{\gamma-1} \frac{u'(r)}{r}. \tag{3.10}$$

Hence (3.10) implies

$$\lim_{r \to R} \frac{p'(r)}{r\rho(r)} = A(q+1)C_\gamma^{\gamma-1} \frac{u'(R)}{R}. \tag{3.11}$$

Fix $\hat{S} \in (S_0, S)$. For any real ℓ, let χ be the real solution of the following initial-value problem:

$$\tilde{\mathcal{L}}\chi = -\ell\chi \quad \text{in } (\hat{S}, S), \tag{3.12}$$

$$\chi(\hat{S}) = 0 \quad \text{and} \quad \chi'(\hat{S}) = 1. \tag{3.13}$$

Denote $\hat{R} = r(\hat{S})$. Now (3.11) implies that there exists $\ell_0 < 0$ such that

$$\ell_0 r\rho(r) - 4p'(r) \le 0 \tag{3.14}$$

in $[\hat{R}, R]$. We claim that $\chi \notin L^2(\hat{S}, S)$ if $\ell \le \ell_0$.

Indeed, if $\chi \in L^2(\hat{S}, S)$, then Lemma 3.1 and (3.8) imply that χ is bounded at S. From (3.12) and (3.13), we obtain

$$\frac{1}{\gamma} \int_{\hat{S}}^{S} r^{-4} p\rho \left(\frac{d\chi}{ds}\right)^2 ds = \ell \int_{\hat{S}}^{S} \chi^2 ds - 4 \int_{\hat{S}}^{S} \frac{p'}{r\rho} \chi^2 ds. \tag{3.15}$$

Now the left-hand side of (3.15) is positive and the right-hand side of (3.15) is nonpositive when $\ell \leq \ell_0$, a contradication. This implies that $\chi \notin L^2(\hat{S}, S)$ for $\ell \leq \ell_0$. Therefore, $\tilde{\mathcal{L}}$ is limit-point type at S.

For (EP), (3.10) and (1.7) imply that

$$\lim_{r \to 0^+} \frac{p'(r)}{r\rho(r)} = -\frac{A}{3}(q+1)C_\gamma^{\gamma-1}u^q(0). \tag{3.16}$$

Now using (3.11) and (3.16), we can choose $\ell_0 < 0$ such that (3.14) holds in $(0, \hat{R})$. Let $\psi(r) = \chi(s(r))$; then Lemma 3.1 implies either

$$\psi(0) \neq 0 \tag{3.17}$$

or

$$|\psi(r)| \leq Cr^3 \quad \text{and} \quad |\psi'(r)| \leq Cr^2 \tag{3.18}$$

for some $C > 0$. Now we can rule out the possibility of (3.18) when $\ell \leq \ell_0$. Indeed, if (3.18) holds, then

$$0 < \frac{1}{\gamma}\int_{-\infty}^{\hat{S}} r^{-4}p\rho \left(\frac{d\chi}{ds}\right)^2 ds = \int_{-\infty}^{\hat{S}} \left(\ell - \frac{4p'}{r\rho}\right)\chi^2 ds < 0,$$

a contradication.

Hence we must have (3.17) when $\ell \leq \ell_0$, i.e., $\chi \notin L^2(-\infty, \hat{S})$. Therefore, $\tilde{\mathcal{L}}$ is a limit-point-type at $-\infty$. The proof is complete. □

An immediate consequence of Lemma 3.2 is that $\tilde{\mathcal{L}}$ is self-adjoint. Indeed, we have the following result (for the proof, see [2]).

Corollary 3.3. *For* (EC) *and* (EPC)*, if $R < \infty$, let $\mathcal{D}_1$ be the set of all functions χ such that*

(i) *χ is differentiable and χ' is absolutely continuous on $[0, \hat{S}]$ for any $\hat{S} < S$,*

(ii) *χ and $\tilde{\mathcal{L}}\chi \in L^2(0, S)$, and*

(iii) *$\chi(0) = 0$.*

Then $\tilde{\mathcal{L}}$ is self-adjoint, i.e.,

$$(\tilde{\mathcal{L}}\chi, \hat{\chi}) = (\chi, \tilde{\mathcal{L}}\hat{\chi}) \tag{3.19}$$

for all χ and $\hat{\chi}$ in $\mathcal{D}_1$.

Similarly, for (EP)*, let $\mathcal{D}_0$ be the set of all functions χ such that*

(i)′ *χ is differentiable and χ' is absolutely continuous over $(-\infty, \hat{S}]$ for any $\hat{S} \in (-\infty, S)$ and*

(ii)′ *χ and $\tilde{\mathcal{L}}\chi \in L^2(-\infty, S)$.*

Then $\tilde{\mathcal{L}}$ is self-adjoint.

Furthermore, using Friedrichs' criteria, we can prove that $\tilde{\mathcal{L}}$ has only a discrete spectrum.

Theorem 3.4. *Let u be a stationary solution of* (EC), (EP), *or* (EPC) *with $R < \infty$. The spectra of $\tilde{\mathcal{L}}$ consist of sequences of strictly increasing eigenvalues $\{\ell_j\}_{j=1}$ with associated eigenfunctions $\{\chi_j\}_{j=1}^\infty$ in $\mathcal{D}_1$ (or $\mathcal{D}_0$).*

Proof. We first claim that no continuous spectrum comes out of S. Indeed, using (3.5), it can be verified that

$$s(r) = S - c_1(R-r)^{q+1} + o((R-r)^{q+1}) \tag{3.20}$$

for r close to R, where $c_1 > 0$ depends on R, $u'(R)$, and γ.

Let

$$a(s) = \frac{1}{\gamma} r^{-4} p(r)\rho(r), \qquad b(s) = \frac{4p'}{r\rho}, \quad \text{and} \quad c(s) = 1.$$

(3.20) then implies

$$a(s) = c_2(S-s)^{2-\varepsilon} + o((S-s)^{2-\varepsilon}), \tag{3.21}$$

where $\varepsilon = \frac{1}{q+1}$ and $c_2 > 0$. Let

$$h(s) = \int_0^s \frac{1}{a(\tau)} d\tau.$$

Then (3.21) implies

$$h(s) = c_3(S-s)^{\varepsilon-1} + o((S-s)^{\varepsilon-1}) \tag{3.22}$$

for s close to S, where $c_3 > 0$.

Hence (3.21) and (3.22) imply that

$$4ah^2 = c_4(S-s)^{\varepsilon} + o((S-s)^{\varepsilon}), \tag{3.23}$$

where $c_4 > 0$. (3.11) now implies that $b(s)$ is bounded at S. Therefore, (3.23) implies that

$$Z(s) = \frac{1}{c}\left\{ b + \frac{1}{4ah^2} \right\} \to +\infty \tag{3.24}$$

as $s \to S$. By Proposition B.3 in Appendix B, no continuous spectrum comes out of S, and $\tilde{\mathcal{L}}$ is totally descrete in R^1. In particular, for (EC) and (EPC), the spectrum of $\tilde{\mathcal{L}}$ is a sequence of eigenvalues $\{\ell_j\}_{j=1}^{\infty}$ such that

$$\lim_{j\to\infty} \ell_j = +\infty. \tag{3.25}$$

For (EP), we also need to prove that (3.24) holds when $s \to -\infty$. From (3.5), we have

$$s = -c_0 r^{-1} + o(r^{-1}) \tag{3.26}$$

for $r \to 0^+$, where $c_0 > 0$. Therefore,

$$a(s) = c_5 s^4 + o(s^4) \tag{3.27}$$

as $s \to -\infty$, where $c_5 > 0$. Let

$$h(s) = \int_s^0 \frac{1}{a(\tau)} d\tau.$$

Then (3.27) implies

$$h(s) = c_6(-s)^{-3} + o(s^{-3}) \tag{3.28}$$

as $s \to -\infty$, where $c_6 > 0$. Hence (3.27) and (3.28) imply

$$4ah^2 = c_7 s^{-2} + o(s^{-2})$$

as $s \to -\infty$. (3.24) then follows from last equation and (3.16). Hence no continuous spectrum comes out from $-\infty$ for (EP). The proof is complete. $\square$

From Lemma 3.4, $\tilde{\mathcal{L}}$ has only the real eigenvalue ℓ. Therefore, λ is either real or purely imaginary for any eigenvalue λ.

From these observations, we then introduce the following notion of stability.

Definition 3.5. *Let u be a ball-type stationary solution of* (EC), (EP), *or* (EPC), *and let* $\{\ell_j\}_{j=1}^{\infty}$ *be the associated eigenvalues of* $\tilde{\mathcal{L}}$ *given in Theorem* 3.4. *u is then called* neutrally stable *if* $\ell_1 > 0$ (*i.e.,* $\lambda_1 = \pm i\sqrt{\ell_1}$ *is purely imaginary*), *is called* unstable *if* $\ell_1 < 0$ (*i.e.,* $\lambda_1 = \pm\sqrt{|\ell_1|}$ *is real*), *and is called* marginally stable *if* $\ell_1 = 0$.

A similar definition can also be given for ground-state- and singularity-type solutions.

Remark 3.6. From Lemma 3.1 and Theorem 3.4, if ψ is an eigenfunction, then $\psi(R)$ is bounded. Furthermore, for (EP), $\psi(r) = O(r^3)$ as $r \to 0^+$. Moreover, the least eigenvalue ℓ_1 can be obtained by a variational method; see, e.g., [3]. Indeed, for (EC) or (EPC), we have

$$\ell_1 = \inf\left\{\frac{Q(\psi)}{I(\psi)} : \psi(1) = 0 \quad \textit{and} \quad \psi \in C^1[1,R]\right\},$$

where

$$Q(\psi) = \int_1^R \left\{r^{-2}p(r)\psi'^2(r) + \frac{4}{\gamma}r^{-3}p'(r)\psi^2(r)\right\}dr$$

and

$$I(\psi) = \frac{1}{\gamma}\int_1^R r^{-2}\rho(r)\psi^2(r)dr.$$

A similar formulation also holds for (EP) with $\psi(r) = O(r^3)$ as $r \to 0^+$.

The following comparison lemma is very useful for testing the stability of stationary solutions.

Lemma 3.7. *Let u be a BC-type stationary solution for* (EC) *or* (EPC). *Then the following hold:*

(i) *If there exists a* $\tilde{\psi} \in C^2([1,R])$ *with* $\tilde{\psi}(1) = 0$, $\tilde{\psi} > 0$ *in* $(1,R]$, *that satisfies*

$$L_0\tilde{\psi} \le 0 \ (\textit{but not} \equiv) \quad \textit{in } (1,R),$$

then u is neutrally stable.

(ii) *If there exists a* $\overline{\psi} \in C^2([1,R])$ *with* $\overline{\psi}(1) = 0$, $\overline{\psi} > 0$ *in* $(1,R]$, *that satisfies*

$$L_0\overline{\psi} \ge 0 \ (\textit{but not} \equiv) \quad \textit{in } (1,R),$$

then u is unstable.

A similar result also holds for a B-type stationary solution u for (EP) *provided the comparison function* $\tilde{\psi}$ (*or* $\overline{\psi}$) *satisfies* $\tilde{\psi}$ (*or* $\overline{\psi}$) $\in L^2_w(0,R)$ *and* $\mathcal{L}\tilde{\psi}$ (*or* $\mathcal{L}\overline{\psi}$) $\in L^2_w(0,R)$.

Proof. Let $\psi_1 > 0$ in (R_0,R) be the associated eigenfunction with respect to ℓ_1 in (3.20). If there is a $\tilde{\psi}$ that satisfies all conditions in (i), then it is easy to see that $\mathcal{L}\tilde{\psi} \le 0$ in $(1,R)$, which implies that

$$0 = \int_1^R (\tilde{\psi}\mathcal{L}\psi_1 - \psi_1\mathcal{L}\tilde{\psi})dr > -\ell_1\int_1^R W\tilde{\psi}\psi_1 dr.$$

Therefore, $\ell_1 > 0$. This proves (i). (ii) and the cases for (EP) can be proved analogously. The proof is complete. □

4. Stability results. In this section, we shall use the methods developed in the preceding section to study the stability of various stationary solutions. We begin with ball-type solutions, proceed to ground-state solutions, and finally conclude with singular solutions.

4.1. Ball-type solutions. We first introduce an auxiliary operator $\tilde{L}$, defined as

$$\tilde{L}\psi \equiv \psi'' - \frac{(q+3)}{r}\psi' + \frac{4q}{r^2}\psi. \tag{4.1}$$

$\tilde{L}$ is closed related to L_0, as can be seen from the following:

$$L_0\psi = \tilde{L}\psi + \left\{(1+q)\psi' - \frac{4q}{r}\psi\right\}\left(\frac{u'}{u} + \frac{1}{r}\right). \tag{4.2}$$

The following results for operator $\tilde{L}$ are very useful in constructing the comparison functions $\tilde{\psi}$ and $\overline{\psi}$ according to Lemma 3.7.

Lemma 4.1. *For any $q > 1$, we have $\tilde{L}(r^4) = 0$ and $\tilde{L}(r^q) = 0$. Moreover, if $q = 4$, we also have $\tilde{L}(r^4 \log r) = 0$.*

Furthermore, if we let (i) $\tilde{\psi} = r^4 - r^q$ *if $q \in (1, 4)$,* (ii) $\tilde{\psi} = r^4 \log r$ *if $q = 4$, and* (iii) $\tilde{\psi} = r^q - r^4$ *if $q \in (4, \infty)$, then we have* (a) $\tilde{L}\tilde{\psi} = 0$ *for $r > 1$,* (b) $\tilde{\psi}(1) = 0$*, and* (c) *the following:*

$$(1+q)\tilde{\psi}' - \frac{4q}{r}\tilde{\psi} > 0 \quad \textit{for} \quad r > 1. \tag{4.3}$$

Proof. The computations are straightforward, so we verify only the last inequality and omit the others. Indeed, for $q \neq 4$,

$$(1+q)(r^q - r^4)' - \frac{4q}{r}(r^q - r^4) = q(q-3)r^{q-1} - 4r^3,$$

and for $q = 4$,

$$(1+q)(r^4 \log r)' - \frac{4q}{r}(r^4 \log r) = 4r^3 \log r + 5r^3.$$

The result follows. □

Next, it is easy to verify the following lemma, so we omit the proof.

Lemma 4.2. *If $u > 0$ in $(1, R)$ and satisfies the equation*

$$u'' + \frac{2}{r}u' + f(u) = 0 \quad \textit{for } r > 1,$$

then

$$\frac{d}{dr}\left(\frac{u'}{u} + \frac{1}{r}\right) = -\frac{1}{r^2u^2}\{(ru' + u)^2 + r^2uf(u)\}.$$

In particular, if $u(\cdot, \alpha, \mu)$ is a solution of (1.6) or (1.8), then $\alpha \leq \mu$ implies

$$\left(\frac{u'}{u} + \frac{1}{r}\right) < 0 \quad \text{in } (1, R). \tag{4.4}$$

We can now establish the stability results for (EC) and (EPC) when $\alpha \leq \mu$.

THEOREM 4.3.

(i) *For any $q > 1$, $\mu > 0$, and $R > 1$, the solution $u_{R,\mu}$ of* (EC) *is neutrally stable.*

(ii) *For any $q > 1$, let $u(\cdot,\alpha,\mu)$ be the solution of* (EPC)*. Then $u(\cdot,\alpha,\mu)$ is neutrally stable if $\alpha \le \mu$.*

Proof. It is not difficult to verify $R(\alpha,\mu) < \infty$ when $\alpha \le \mu$ in (ii). Let $\tilde{\psi}$ be given as in Lemma 4.1. Then for both (i) and (ii), Lemmas 4.1 and 4.2 imply $L\tilde{\psi} < 0$ in $(1, R)$.

Thus Lemma 3.7 implies that $u_{R,\mu}$ and $u(\cdot,\alpha,\mu)$ with $0 < \alpha \le \mu$ are neutrally stable. The proof is complete. □

We can also establish other stability results for (EPC) by choosing appropriate comparison functions and applying Lemma 3.7. For example, we can prove the following theorem.

THEOREM 4.4. *For* (EPC)*, we have the following:*

(i) *If $q \in (1, 3]$, then all BC-type solutions are neutrally stable.*

(ii) *For any $q > 1$, there is $R_q > 1$ such that u is neutrally stable whenever the first zero R of u is less than R_q.*

Proof. (i) It is known that $R(\alpha,\mu) < \infty$ for any $\alpha > 0$ and $\mu > 0$ when $q \in (1, 3]$; see, e.g., [18]. Let $\tilde{\psi} = r^3 - 1$. Then $\tilde{\psi}(1) = 0$, $\tilde{\psi} > 0$ in $(0,\infty)$, and

$$L_0\tilde{\psi} = \left\{ \frac{u'}{u}(3-q)r^2 + \frac{4q}{r} \right\},$$

which is negative in $(0,\infty)$ if $q \in (1, 3]$. Thus by Lemma 3.7(i), u is neutrally stable.

(ii) Let $\tilde{\psi} = \log r$. Then $\tilde{\psi}(1) = 0$, $\tilde{\psi} > 0$ in $(1,\infty)$, and

$$L_0\tilde{\psi} = -3r^{-2} + \frac{1}{r}\frac{u'}{u}\{(1+q) - 4q\log r\}.$$

Therefore, $L_0\tilde{\psi} < 0$ in $(1, R)$ if $R \le R_q \equiv \exp(\frac{1+q}{4q})$. The result also follows from Lemma 3.7(i). The proof is complete. □

Remark 4.5. By picking a comparison function $\tilde{\psi}$ different from $\log r$ in Theorem 4.4(ii), we can also obtain another $\tilde{R}_q$, which ensures that u is neutrally stable when $R \le \tilde{R}_q$.

By choosing an appropriate comparison function, we obtain the following stability results for (EP).

THEOREM 4.6. *For* (EP)*, we have the following:*

(i) *If $q \in (1, 3)$, then any B-type solution is neutrally stable.*

(ii) *If $q = 3$, then any B-type solution is marginally stable.*

(iii) *If $q \in (3, 5)$, then any B-type solution is unstable.*

Proof. Let $\tilde{\psi} = r^3$. Then $\tilde{\psi}(0) = \tilde{\psi}'(0) = 0$ and $\tilde{\psi} > 0$ in $(0,\infty)$. Furthermore, we have

$$L_0\tilde{\psi} = (3-q)r^2\frac{u'}{u}.$$

Hence the result follows by Lemma 3.7. The proof is complete. □

Proof of Theorem 1.1. Combining the results from Theorems 4.3, 4.4, and 4.6, we obtain Theorem 1.1. □

4.2. Ground-state solutions. From section 2, we know that if a ground-state-type solution u has a finite total mass, then it is necessary that u have a fast decay rate, i.e.,

$$\lim_{r\to\infty} ru(r) = m \in (0,\infty). \tag{4.5}$$

In this section, we will prove that the linearized operator $\mathcal{L}$ associated with u has a continuous spectrum $(0,\infty)$. Therefore, u cannot be neutrally stable. In fact, it is either marginally stable or unstable.

LEMMA 4.7. *If u is a G- or GC-type solution and satisfies* (4.5)*, then the linearized operator $\mathcal{L}$ of u is discrete below* 0 *and has a continuous spectrum* $(0,\infty)$*.*

Proof. In Lemma 3.4, we have shown that no continuous spectrum comes from $r=0$ for $\mathcal{L}$ in (EP). Therefore, we need only study $\mathcal{L}$ as $r\to\infty$. We may assume that $m=1$ in (4.5). We then have

$$p(r) = \tilde{A}r^{-1-q} + o(r^{-1-q})$$

and

$$p'(r) = -(1+q)\tilde{A}r^{-q} + o(r^{-q})$$

as $r\to\infty$, where $\tilde{A} = AC_\gamma^\gamma$. As before, we have the following asymptotic expansions for the coefficients of $\mathcal{L}$ as $r\to\infty$:

$$a(r) = r^{-2}p(r) = \tilde{A}r^{-3-q} + o(r^{-3-q}),$$

$$b(r) = \frac{4}{\gamma}r^{-3}p'(r) = -4q\tilde{A}r^{-5-q} + o(r^{-5-q}),$$

and

$$c(r) = \frac{1}{\gamma}r^{-2}\rho(r) = \hat{A}r^{-2-q} + o(r^{-2-q})$$

as $r\to\infty$, where $\hat{A}>0$ is a constant. Therefore, for large fixed $\hat{r}$, we have

$$h(r) = \int_{\hat{r}}^{r} \frac{d\tau}{a(\tau)} = \{\tilde{A}(4+q)\}^{-1}r^{4+q} + o(r^{4+q})$$

as $r\to\infty$.

We claim that

$$Z(r) = \frac{1}{c(r)}\left\{b(r) + \frac{1}{4a(r)h^2(r)}\right\} \to 0 \quad \text{as } r\to\infty. \tag{4.6}$$

Indeed, it is clear that

$$4a(r)h^2(r) = 4\tilde{A}^{-1}(4+q)^{-2}r^{q+5} + o(r^{5+q}) \quad \text{as } r\to\infty.$$

Therefore, we have

$$\begin{aligned} b(r) + \frac{1}{4a(r)h^2(r)} &= \tilde{A}\left\{\frac{(4+q)^2}{4} - 4q\right\}r^{-5-q} + o(r^{-5-q}) \\ &= \frac{\tilde{A}}{4}(q-4)^2r^{-5-q} + o(r^{-5-q}). \end{aligned}$$

Hence

$$Z(r) = A^* r^{-3}(q-4)^2 + o(r^{-3})$$

for some $A^* > 0$. (4.6) follows. Now by Proposition B.3 (II)–(III) in Appendix B, the linearized operator $\mathcal{L}$ of u has a continuous spectrum $(0,\infty)$ and is descrete below 0. The proof is complete. □

An immediate consequence of Lemma 4.7 is the following theorem for ground-state-type stationary solutions.

THEOREM 4.8. *Any ground-state-type solution of* (EC), (EP), *or* (EPC) *is either marginally stable or unstable.*

4.3. Singular solutions. In this section, we will continuously apply Friedrichs' criteria to study the stability of singularity-type solutions. We know that if $q \in (3,5)$ and u is a singular solution of (EP) with finite total mass, then u has a weak singularity at $r=0$, i.e.,

$$\lim_{r\to o^+} r^\sigma u(r) = m \in (0,\infty). \tag{4.7}$$

As in section 4.2, we are interested in the limit of $Z(r)$ as $r \to 0^+$. (4.7) now implies the following expansions:

$$a(r) = r^{-2}p(r) = \tilde{A} r^{-\sigma(q+1)-2} + o(r^{-\sigma(q+1)-2}),$$

$$b(r) = \frac{4}{\gamma} r^{-3} p'(r) = -4\sigma q \tilde{A} r^{-\sigma(q+1)-4} + o(r^{-\sigma(q+1)-4}),$$

and

$$c(r) = \hat{A} r^{-2-\sigma q} + o(r^{-2-\sigma q})$$

as $r \to 0^+$ for some positive constants $\tilde{A}$ and $\hat{A}$.
Therefore,

$$h(r) = \int_0^r \frac{ds}{a(s)} = \tilde{A}^{-1}\{3+\sigma(q+1)\}^{-1} r^{\sigma(q+1)+3},$$

with $h(0)=0$.

It is straightfoward to compute

$$\begin{aligned} Z(r) &= A^* r^{-\sigma-2}\left\{\frac{1}{4}\left[3+\sigma(q+1)\right]^2 - 4\sigma q\right\} + o(r^{-\sigma-2}) \\ &= \frac{A^*}{4} r^{-\sigma-2}\{4\sigma^2+4\sigma-7\} + o(r^{-\sigma-2}) \end{aligned}$$

for some positive constant A^*.

Hence we obtain the following lemma.

LEMMA 4.9. *Let u be a singular solution of* (EP) *satisfying* (4.7). *Then we have the following:*

(i) *if* $q \in (3, q^+)$, *then* $\lim_{r\to 0^+} Z(r) = +\infty$;

(ii) *if* $q = q^+$, *then* $\lim_{r\to 0^+} Z(r) = 0$;

(iii) *if* $q \in (q^+, 5)$, *then* $\lim_{r\to 0^+} Z(r) = -\infty$;

and

$$\Omega_0 = \int_0^{\hat{r}} \left\{ \frac{c(r)}{-a(r)Z(r)} \right\}^{\frac{1}{2}} dr < \infty.$$

Therefore, by applying Theorem 4.6(iii), Lemma 4.9, and Proposition B.3, we obtain the following theorem for singularity-type solutions.

THEOREM 4.10. *For problem* (EP)*, we have the following:*

(i) *If $q \in (3, q^+)$, then any SB-type solution is unstable and has no continuous spectrum. Any SG-type solution is also unstable but has a continuous spectrum* $(0, \infty)$.

(ii) *If $q = q^+$, then any SB-type and SG-type solution is unstable and has a continuous spectrum* $(0, \infty)$.

(iii) *If $q \in (q^+, 5)$, then any SB-type and SG-type solution is unstable, and there is a sequence of pure imaginary eigenvalues $\{\lambda_k\}$ such that* $\lim_{k\to\infty} \lambda_k^2 = -\infty$.

Proof. For any $q \in (3,5)$ and for an SB-type solution u, choose $\tilde{\psi} = r^3$. Then we have

$$L\tilde{\psi} = (3-q)r^2 \frac{u'}{u} > 0.$$

Therefore, by modifying the proof of Lemma 3.5, we can prove that u is unstable. The remaining results follow from Lemma 4.9 and Proposition B.3. The details of the proof are omitted and the proof is complete. □

5. Effects of viscosity. In this section, we shall study the effect of viscosity on the stability problem of stationary solutions. From equation (1.2), it is clear that stationary solutions for inviscid flow are also solutions for viscous flow. As we have seen in the previous sections, the best possibilities for stationary solutions are neutrally stable in the inviscid case. It is known that neutral stability is very sensitive to disturbances. Therefore, we need to know what effect viscosity has on neutrally stable stationary solutions.

Since the gaseous mass is not confined from outside, its outer surface is a free surface maintained by the attraction of the core and its own gravitational forces. Presumably, the surface of the gas should be very sensitive to a direct disturbance of it. In this section, we show that this is the case, as mentioned in Theorem 1.2.

When viscosity its present, the linearized equation is

$$\mathcal{L}\psi = \lambda^2 W\psi - \lambda\nu\hat{\mathcal{L}}\psi, \tag{5.1}$$

where

$$\hat{\mathcal{L}}\psi \equiv \frac{1}{\gamma}(r^{-2}\psi')' \tag{5.2}$$

or, equivalently,

$$\left\{ r^{-2}\left(p(r) + \lambda\frac{\nu}{\gamma} \right)\psi' \right\}' - \frac{1}{\gamma}\{4r^{-3}p'(r) + \lambda^2 r^{-2}\}\psi = 0. \tag{5.3}$$

When $\nu > 0$, the eigenvlaue equation (5.1) is linear for ψ but quadratic for λ, which is different from ordinary eigenvalue problems. Indeed, if $\nu = 0$ in (5.1), then (5.1) is linear for $\ell = -\lambda^2$. Since the coefficients of $\mathcal{L}$, $\hat{\mathcal{L}}$, and W are real, it is easy to see that if $\{\lambda, \psi\}$ is a solution of (5.1), then its conjugate $\{\overline{\lambda}, \overline{\psi}\}$ is also a solution.

This property does not affect the stability, which depends on the sign of Reλ in the stationary solution.

In this section, we concentrate on the effects of viscosity and boundary disturbances. Therefore, we restrict our study to ball-type solutions which are neutrally stable. The problems of unstable stationary solutions, ground-state solutions, and singularity-type solutions will be left for future study.

We first consider (EC) and (EPC) and then continue by studying (EP).

Let

$$\lambda^* = \frac{\gamma}{\nu} P(R_0). \tag{5.4}$$

When $R_0 = 1$, we will prove that (5.1) is regular on $[1, R]$ when $\lambda \notin [-\lambda^*, 0]$. Indeed, for $\lambda \neq -\lambda^*$, let $\psi(\cdot, \lambda) = \psi(\cdot, \lambda, \nu)$ be the solution of (5.3) that satisfies the initial conditions

$$\psi(1, 0) = 0 \tag{5.5}$$

and

$$\psi'(1, \lambda) = 1. \tag{5.6}$$

We can then prove the following result.

Lemma 5.1. *Let u be a BC-type stationary solution of* (EC) *or* (EPC). *If $\lambda \notin [-\lambda^*, 0]$, then $\psi(\cdot, \lambda)$ is C^2 on $[1, R]$ and is analytic in $\lambda \in \mathbf{C} - [-\lambda^*, 0]$. Furthermore, if $\lambda \in (-\lambda^*, 0)$, then either $\psi(\cdot, \lambda)$ is bounded at $r = \hat{r}$ or $|\psi(r, \lambda)|$ grows like $|\log |r - \hat{r}||$ as $r \to \hat{r}$, where $\hat{r} \in (1, R)$ satisfies $p(\hat{r}) + \lambda \frac{\nu}{\gamma} = 0$. If $\lambda = -\lambda^*$, then any nontrivial solution ψ of* (5.3) *is unbounded in a neighborhood of $r = 1$. The case in which $\lambda = 0$ was studied in Lemma* 3.1.

Proof. Let $\lambda = \lambda_1 + i\lambda_2$ and $\psi = \psi_1 + i\psi_2$ in (5.3) and denote

$$a = r^{-2}\left(p + \lambda_1 \tfrac{\nu}{\gamma}\right), \qquad b = \lambda_2 \frac{\nu}{\gamma} r^{-2},$$

$$c = \frac{1}{\gamma}\{4r^{-3}p' + r^{-2}(\lambda_1^2 - \lambda_2^2)\}, \qquad d = \frac{2}{\gamma}\lambda_1\lambda_2 r^{-2}.$$

Then it is clear that $a^2(\hat{r}) + b^2 = 0$ for some $\hat{r} \in [1, R]$ if and only if $\lambda_2 = 0$ and $\lambda_1 \in [-\lambda^*, 0]$. In this case, $p(\hat{r}) + \lambda_1 \frac{\nu}{\gamma} = 0$.

Now (5.3) can be written as the following system of equations:

$$\begin{aligned} (a\psi_1' - b\psi_2')' &= c\psi_1 - d\psi_2, \\ (b\psi_1' + a\psi_2')' &= d\psi_1 + c\psi_2. \end{aligned} \tag{5.7}$$

For $\lambda \notin [-\lambda^*, 0]$, denote

$$\tilde{\psi}_1 = a\psi_1 - b\psi_2 \quad \text{and} \quad \tilde{\psi}_2 = b\psi_1 + a\psi_2. \tag{5.8}$$

We then have

$$\begin{aligned} \psi_1 &= (a\tilde{\psi}_1 + b\tilde{\psi}_2)(a^2 + b^2)^{-1}, \\ \psi_2 &= (-b\tilde{\psi}_1 + a\tilde{\psi}_2)(a^2 + b^2)^{-1}. \end{aligned} \tag{5.9}$$

By a straightforward but lengthy computation on (5.7), we obtain the following system of equations for $\tilde{\psi}_1$ and $\tilde{\psi}_2$:

$$\begin{aligned}\tilde{\psi}_1'' &= \tilde{A}\tilde{\psi}_1' + \tilde{B}\tilde{\psi}_1 + \tilde{C}\tilde{\psi}_2' + \tilde{D}\tilde{\psi}_2,\\ \tilde{\psi}_2'' &= \tilde{A}\tilde{\psi}_2' + \tilde{B}\tilde{\psi}_2 - \tilde{C}\tilde{\psi}_1' - \tilde{D}\tilde{\psi}_2,\end{aligned} \tag{5.10}$$

where

$$\tilde{a} = a(a^2+b^2)^{-1}, \qquad \tilde{b} = b(a^2+b^2)^{-1}, \tag{5.11}$$

and

$$\begin{aligned}\tilde{A} &= a'\tilde{a} + b'\tilde{b},\\ \tilde{B} &= (a''+c)\tilde{a} + (b''+d)\tilde{b} + a'\tilde{a}' + b'\tilde{b}',\\ \tilde{C} &= a'\tilde{b} - b'\tilde{a},\\ \tilde{D} &= (a''+c)\tilde{b} - (b''+d)\tilde{s} + a'\tilde{b}' - b'\tilde{a}'.\end{aligned} \tag{5.12}$$

Since the coefficients of (5.10) are continuous on $[1,R]$, then $\tilde{\psi}_1'$ and $\tilde{\psi}_2'$ are C^2 on $[1,R]$ and analytic in $\lambda \in \mathbf{C} - [-\lambda^*, 0]$. Hence ψ_1 and ψ_2 have the same properties as $\tilde{\psi}_1$ and $\tilde{\psi}_2$. This proves the first part of the lemma.

To study $\lambda \in (-\lambda^*, 0)$, we write (5.3) as

$$\psi'' + \left\{\frac{1}{r-\hat{r}} + g(r)\right\}\psi' + \frac{1}{r-\hat{r}}f(r)\psi = 0 \quad \text{for } r < \hat{r},$$

where g and f are analytic at $\hat{r}$. Hence $\hat{r}$ is a regular singular point. Therefore, by a standard theorem (see, e.g., [2]), this implies that ψ either is bounded at $\hat{r}$ or grows logarithmically at $\hat{r}$.

Finally, if $\lambda = -\lambda^*$, then $p(1) = \lambda^* \frac{\nu}{\gamma}$. Let $s = r-1$; then (5.3) can be written as

$$\psi'' + \left(\frac{2}{s} + g\right)\psi' + \left(\frac{c_2}{s^2} + \frac{c_1}{s} + f\right)\psi = 0 \qquad \text{for } s > 0,$$

where g and f are continuous at $s = 0$ and $c_2 > 0$.

Let

$$\mu_1 = \frac{1}{2}(-1 + \sqrt{1-4c_2}) \qquad \text{and} \qquad \mu_2 = \frac{1}{2}(-1 - \sqrt{1-4c_2}).$$

If $\mu_1 \neq \mu_2$, then ψ behaves asymptotially like s^{μ_1} or s^{μ_2} as $s \to 0^+$. If $\mu_1 = \mu_2 = -\frac{1}{2}$, then $|\psi(s)|$ behaves asymptotically like $s^{-\frac{1}{2}}$ or $s^{-\frac{1}{2}}|\log s|$ as $s \to 0^+$. In any case, ψ is unbounded at $s = 0$. The case in which $\lambda = 0$ was studied in Lemma 3.1. The proof is complete. □

Considering (1.13) and Lemma 5.1, we introduce the following notion.

DEFINITION 5.2. *For* $\lambda \notin [-\lambda^*, 0], \psi(\cdot, \lambda)$ *is called a* stable mode *if* $\text{Re}\lambda < 0$*, an* unstable mode *if* $\text{Re}\lambda > 0$*, and a* marginally stable mode *if* $\text{Re}\lambda = 0$.

In the following, we shall study the relationship between the sign of $\text{Re}\lambda$ and $\psi(R, \lambda)$, i.e., how the disturbance of the gas surface influences the stability of the stationary solution u.

Since $\psi(\cdot, \lambda)$ is C^2 in $[1,R]$ for $\lambda \notin [-\lambda^*, 0]$, $\psi(R, \lambda)$ and $\psi'(R, \lambda)$ satisfies homogeneous boundary conditions at R, i.e.,

$$a_j \psi_j'(R) + b_j \psi_j(R) = 0, \tag{5.13}$$

where $\psi_1 = \text{Re}\psi$ and $\psi_2 = \text{Im}\psi$, $a_j = a_j(\lambda)$ and $b_j = b_j(\lambda)$ are analytic in $\lambda \notin [-\lambda^*, 0]$ for $j = 1, 2$.

When $a_j \neq 0$, denote

$$\kappa_j(\lambda) = \frac{b_j(\lambda)}{a_j(\lambda)}. \tag{5.14}$$

Then (5.13) can be written as

$$\frac{\psi_j'(R)}{\psi_j(R)} = -\kappa_j. \tag{5.15}$$

When $a_j = 0$, i.e., ψ_j satisfies the Dirichlet boundary condition $\psi_j(R) = 0$, we adopt the convention $\kappa_j = +\infty$.

We can now introduce the notion of the stability of stationary solutions with respect to the boundary conditions (5.5) and (5.15) (or (5.13)).

Definition 5.3. *Let u be a BC-type stationary solution for* (EC) *or* (EPC). *Then u is called* stable *with respect to* (5.5) *and* (5.15) *if any eigenvalue λ of* (5.1), (5.5), *and* (5.15) *satisfies* $\text{Re}\lambda < 0$. *u is called* unstable *if there is an eigenvalue $\tilde{\lambda}$ of* (5.1), (5.5), *and* (5.15) *such that* $\text{Re}\tilde{\lambda} > 0$. *$u$ is called* marginally stable *if any eigenvalue λ of* (5.1), (5.5), *and* (5.15) *satisfies* $\text{Re}\lambda \leq 0$ *and equality holds for some $\tilde{\lambda}$.*

The stability problem with respect to boundary condition (5.15) can also be studied by making the following observation:

Denote

$$\mathbf{C}^+ = \{\lambda \in \mathbf{C} : \text{Re}\lambda > 0\}, \qquad \mathbf{C}^- = \{\lambda \in \mathbf{C} : \text{Re}\lambda < 0\},$$
$$\text{and} \quad \mathbf{C}^0 = \{\lambda \in \mathbf{C} : \text{Re}\lambda = 0\}.$$

For any stationary solution u and any

$$(\kappa_1, \kappa_2) \in \overline{\mathbf{R}}^2 \equiv \mathbf{R}^2 \cup \{(k_1, \infty) : \kappa_1 \in \mathbf{R}^1\} \cup \{(\infty, k_2) : \kappa_2 \in \mathbf{R}^1\} \cup \{(\infty, \infty)\},$$

denote by $\sigma(\kappa_1, \kappa_2)$ the set of eigenvalues of (5.1), (5.5), and (5.15). Then define

$$K_s = K_s(u) \equiv \{(\kappa_1, \kappa_2) : \sigma(\kappa_1, \kappa_2) \subset \mathbf{C}^-\},$$

$$K_u = K_u(u) \equiv \{(\kappa_1, \kappa_2) : \sigma(\kappa_1, \kappa_2) \cap \mathbf{C}^+ \neq \phi\},$$

and

$$K_m = K_m(u) = \{(\kappa_1, \kappa_2) : \sigma(\kappa_1, \kappa_2) \cap \mathbf{C}^+ = \phi \quad \text{and} \quad \sigma(\kappa_1, \kappa_2) \cap \mathbf{C}^0 \neq \phi\}.$$

From Lemma 5.1, we know that any one of K_s, K_u, and K_m is nonempty. Hence the stability of u with respect to a given (κ_1, κ_2) is equivalent to deciding to which set—K_s, K_u, or K_m—(κ_1, κ_2) belongs. In general, for a given u, it is not easy to completely identify K_s, K_u, and K_m. However, we shall find some subsets of K_s and K_u that will give us sufficient conditions to determine whether u is stable or unstable with respect to given (κ_1, κ_2).

We first prove the following stability result.

THEOREM 5.4. *Let u be a neutrally stable BC-type stationary solution of* (EC) *or* (EPC) *when $\nu = 0$. Then for any $\nu > 0$, we have*

$$\{(\kappa_1, \kappa_2) : \kappa_1 \geq 0, \quad \kappa_2 \geq 0\} \subset K_s(u), \tag{5.16}$$

i.e., u is stable if

$$\kappa_1 \geq 0 \quad \textit{and} \quad \kappa_2 \geq 0 \tag{5.17}$$

or, equivalently, if

$$\psi_j'(R)\psi_j(R) \leq 0 \tag{5.18}$$

for $j = 1, 2$.

Proof. Since u is assumed to be neutrally stable when $\nu = 0$, 0 is not an eigenvalue of (5.1), (5.5), and (5.15). If $\lambda \in (-\lambda^*, 0)$, then there is nothing to prove. Hence we consider the case where $\lambda \notin [-\lambda^*, 0)$ and is an eigenvalue with respect to (5.15) such that $(\kappa_1(\lambda), \kappa_2(\lambda))$ satisfies (5.17). We must prove that

$$\text{Re}\lambda < 0. \tag{5.19}$$

Indeed, multiply (5.1) by $\overline{\psi}$ and then integrate from 1 to R; λ satisfies

$$a\lambda^2 + b\lambda + c = 0, \tag{5.20}$$

where

$$a = \frac{1}{\gamma}\int_1^R r^{-2}\rho(r)(\psi_1^2 + \psi_2^2)dr > 0, \tag{5.21}$$

$$b = -\nu\int_1^R \overline{\psi}\hat{\mathcal{L}}\psi dr, \quad \text{and} \quad c = -\int_1^R \overline{\psi}\mathcal{L}\psi dr.$$

Since u is assumed to be neutrally stable when $\nu = 0$, we have $\ell_1 > 0$ in (3.1). Moreover, ψ is C^2 on $[1, R]$. Hence Remark 3.6 implies that

$$c > 0. \tag{5.22}$$

Now let

$$b = b_1 + ib_2, \tag{5.23}$$

where

$$\begin{aligned} b_1 &= -\nu\int_1^R (\psi_1\hat{\mathcal{L}}\psi_1 + \psi_2\hat{\mathcal{L}}\psi_2) \\ &= \nu\left\{\sum_{j=1}^2 \int_1^R r^{-2}(\psi_j')^2 - \sum_{j=1}^2 R^{-2}\psi_j'(R)\psi_j(R)\right\} \end{aligned}$$

and

$$\begin{aligned} b_2 &= -\nu\int_1^R (\psi_1\hat{\mathcal{L}}\psi_2 - \psi_2\hat{\mathcal{L}}\psi_1) \\ &= \frac{\nu}{\gamma}R^{-2}(\psi_2(R)\psi_1'(R) - \psi_1(R)\psi_2'(R)). \end{aligned}$$

Assumption (5.18) implies that

$$b_1 > 0. \tag{5.24}$$

Now we are going to show that the root λ of (5.20) satisfies (5.19) provided the coefficients satisfy (5.21)–(5.24). It is clear that the roots λ of (5.20) are given by

$$\lambda_{\pm} = \frac{1}{2a}\{-(b_1 + ib_2) \pm (b^2 - 4ac)^{\frac{1}{2}}\}. \tag{5.25}$$

Let

$$X = b_1^2 - b_2^2 - 4ac \quad \text{and} \quad Y = b_1 b_2.$$

Then

$$b^2 - 4ac = X + 2iY.$$

Moreover, if x and y are real numbers such that

$$(x + iy)^2 = X + iY,$$

then

$$x^2 = \frac{1}{2}\{X + (X^2 + 4Y^2)^{\frac{1}{2}}\}.$$

To show (5.19), it suffices to prove that $b_1 > |x|$, i.e.,

$$b_1^2 > x^2. \tag{5.26}$$

By (5.21) and (5.22), we have

$$2b_1^2 - X = b_1^2 + b_2^2 + 4ac > 0. \tag{5.27}$$

It is easy to check that

$$(2b_1^2 - X)^2 - (X^2 + 4Y^2) = 16ac. \tag{5.28}$$

Hence (5.26) follows from (5.21), (5.22), (5.27), and (5.28). The proof is complete. □

Next, we prove the following instability results.

LEMMA 5.5. *Let u be a neutrally stable BC-type stationary solution of* (EC) *or* (EPC) *when $\nu = 0$. For any $\nu > 0$, if λ is real and $\lambda > 0$, we have*

$$\psi(R, \lambda) > 0 \quad \textit{and} \quad \psi'(R, \lambda) > 0. \tag{5.29}$$

Furthermore, we have

$$\lim_{\lambda \to 0^+} \frac{\psi'(R, \lambda)}{\psi(R, \lambda)} = +\infty \tag{5.30}$$

and

$$\lim_{\lambda \to \infty} \frac{\psi'(R, \lambda)}{\psi(R, \lambda)} = +\infty. \tag{5.31}$$

Proof. If $\lambda > 0$, then $p(r) + \lambda\frac{\nu}{\gamma} > 0$ in $[1, R]$. Integrating (5.3) from 1 to r and using (5.5) and (5.6), we obtain

(5.32)
$$r^{-2}\left(p(r) + \lambda\frac{\nu}{\gamma}\right)\psi'(r) = \left(p(1) + \lambda\frac{\nu}{\gamma}\right) + \frac{1}{\gamma}\int_1^r \{4s^{-3}p'(s) + \lambda^2 s^{-2}\}\psi(s,\lambda)ds.$$

If λ^2 is large enough that

(5.33)
$$\lambda^2 + 4s^{-1}p'(s) \geq 0 \quad \text{in } [1, R],$$

then (5.5), (5.6), and (5.32) imply

(5.34)
$$\psi(r,\lambda) > 0 \quad \text{and} \quad \psi'(r,\lambda) > 0 \quad \text{in } [1, R].$$

In particular, (5.29) holds.

Now by applying Theorem 5.4, we claim that (5.29) also holds for any $\lambda > 0$. Otherwise, by the continuous dependence of $\psi(R,\lambda)$ with respect to λ, we have $\psi'(R,\lambda_1) = 0$ or $\psi(R,\lambda_1) = 0$ for some $\lambda_1 > 0$. Since (5.18) is satisfied by this λ_1, Theorem 5.4 implies $\lambda_1 < 0$, a contradiction. Hence (5.29) holds for any $\lambda > 0$.

To show (5.30), we note that u is neutrally stable and by Proposition A.1 in Appendix A, we have

(5.35)
$$\lim_{r\to R^-}(R-r)^q\psi(r,0) = c_0 > 0$$

and

(5.36)
$$\lim_{r\to R^-}(R-r)^{q+1}\psi'(r,0) = c_1 > 0.$$

From (5.35), (5.36), and (5.29), it is not difficult to prove that (5.30) holds. The details of the proof are omitted.

Finally, it remains to prove (5.31). If $\lambda > 0$ and is large enough, then (5.3), (5.33), and (5.34) imply that

(5.37)
$$\psi''(r,\lambda) > 0 \quad \text{in } [1, R].$$

Moreover, by (5.32), there is a positive constant c_2 that is independent on λ such that for a large λ, we have

(5.38)
$$\psi'(R,\lambda) \geq \lambda c_2 \int_r^R \psi(s,\lambda)ds$$

for $r \in [\frac{1}{2}R, R]$. Now for any $s \in [\frac{1}{2}R, R]$, write

(5.39)
$$\psi(s,\lambda) = \psi(R,\lambda) + \psi'(R,\lambda)(s-R) + \frac{1}{2}\psi''(\tilde{r},\lambda)(s-R)^2$$

for some $\tilde{r} \in (s, R)$. Subsituting (5.39) into (5.38) and using (5.37), we obtain

(5.40)
$$\psi'(R,\lambda)\left\{1 + \frac{1}{2}\lambda c_2(R-r)^2\right\} \geq \lambda c_2\psi(R,\lambda)(R-r).$$

If we choose r such that $(R-r)\lambda^{\frac{1}{2}} = 1$, then (5.40) implies that for a large λ, we have

$$\psi'(R,\lambda) \geq c_3\lambda^{\frac{1}{2}}\psi(R,\lambda),$$

where the positive constant c_3 is independent of λ. Hence (5.31) follows. The proof is complete. □

For any real $\lambda \notin [-\lambda^*, 0]$, $\psi(r,\lambda)$ is a real function, i.e., $\psi_2 = \mathrm{Im}\psi \equiv 0$. $\kappa_2(\lambda)$ is then undetermined for all real $\lambda \notin [-\lambda^*, 0]$. However, we can define $\kappa_2(\lambda)$ for these λ by going through the following limiting process.

Let λ_1 and λ_2 be real numbers such that $\lambda_1 \notin [-\lambda^*, 0]$ and $|\lambda_2| \neq 0$ and is sufficiently small. We then have

$$\psi(r, \lambda_1 + i\lambda_2) = \psi(r,\lambda_1) + i\lambda_2\frac{\partial\psi}{\partial\lambda}(r,\lambda_1) + o(|\lambda_2|^2) \tag{5.41}$$

as $\lambda_2 \to 0$. Therefore, for any real $\lambda_1 \notin [-\lambda^*, 0]$, we can define

$$\kappa_2(\lambda_1) = \frac{\partial^2\psi}{\partial r\partial\lambda}(R,\lambda_1)/\frac{\partial\psi}{\partial\lambda}(R,\lambda_1). \tag{5.42}$$

It is not difficult to prove that $\kappa_2(\lambda)$ is well defined and is continuous for $\lambda \in \mathbf{C} - [-\lambda^*, 0]$.

Now by applying Lemma 5.5, we have that following instability result.

Theorem 5.6. *Let u be a neutrally stable BC-type stationary solution of* (EC) *or* (EPC) *when $\nu = 0$. Then for any $\nu > 0$, there is a positive constant $\kappa^* = \kappa^*(\nu, u)$ such that for any $\kappa_1 < -\kappa^*$, there is a nonempty open set $U(\kappa_1, \nu, u)$ such that u is unstable with respect to* (5.5) *and* (5.15) *for (κ_1, κ_2) with $\kappa_2 \in U(\kappa_1, \nu, u)$.*

Proof. For any $\nu > 0$, let

$$\kappa^*(\nu, u) = \min\left\{\frac{\psi'(R,\lambda,\nu)}{\psi(R,\lambda,\nu)} : \lambda \in (0,\infty)\right\}.$$

By (5.30) and (5.31), we have $\kappa^*(\nu, u) > 0$. If $\kappa_1 < -\kappa^*$, then there is $\lambda_1 > 0$ such that

$$\frac{\psi'(R,\lambda_1)}{\psi(R,\lambda_1)} = -\kappa_1.$$

Let

$$U(\kappa_1,\nu,u) = \{\kappa_2 \in (-\infty,\infty] : (\kappa_1,\kappa_2) \in K_u\}.$$

Then (5.42) implies that

$$\kappa_2(\lambda_1) \in U(\kappa_1,\nu,u).$$

Thus $U(\kappa_1,\nu,u)$ is nonempty. It is clear that $U(\kappa_1,\nu,u)$ is open, and the result follows. The proof is complete. □

Proof of Theorem 1.2. Theorem 1.2 follows from Theorems 5.4 and 5.6. □

We now come to (EP). In this case, (5.3) has a singularity at $r = 0$ even for $\lambda \notin [-\lambda^*, 0)$. Therefore, we need to modify our argument to obtain a result as in Lemma 5.1. Indeed, the initial conditions (5.5) and (5.6) will be replaced with

$$r^{-2}\psi(r,\lambda) = 0 \quad \text{at } r = 0 \tag{5.43}$$

and

$$(r^{-2}\psi(r,\lambda))' = 1 \quad \text{at } r = 0. \tag{5.44}$$

We then have the following result.

LEMMA 5.7. *Let u be a B-type solution of* (EP). *Then the solution $\psi(r,\lambda)$ of* (5.3), (5.43), *and* (5.44) *exists in a neighborhood of $r = 0$ if $\lambda \neq -\lambda^* \equiv \frac{\gamma}{\nu}p(0)$. Furthermore, $\psi(\cdot,\lambda)$ has the same property as in Lemma* 5.1.

Proof. Following the same argument as in the proof of Lemma 5.1, we have equation (5.9) for $\tilde{\psi}_1$ and $\tilde{\psi}_2$ in $r > 0$. For $\lambda \neq -\lambda^*$, after a careful computation, (5.9) can be written as

$$\tilde{\psi}_1'' = \left(-\frac{2}{r} + g_1\right)\tilde{\psi}_1' + \left(\frac{2}{r^2} + f_1\right)\tilde{\psi}_1 + g_2\tilde{\psi}_2' + f_2\tilde{\psi}_2 \tag{5.45}$$

and

$$\tilde{\psi}_2'' = \left(-\frac{2}{r} + g_1\right)\tilde{\psi}_2' + \left(\frac{2}{r^2} + f_1\right)\tilde{\psi}_2 - g_2\tilde{\psi}_1' - f_2\tilde{\psi}_1, \tag{5.46}$$

where $g_j(r)$ and $rf_j(r)$ are continuous (in fact, C^2) at $r = 0$ for $j = 1, 2$. Now $r = 0$ is a regular singular point in (5.45) and (5.46). By a standard argument, we can prove that there is bounded solution $\{\tilde{\psi}_1, \tilde{\psi}_2\}$ of (5.45) and (5.46). Moreover, they satisfy

$$\begin{aligned} \tilde{\psi}_1(r) &= a_0 r + o(r), \\ \tilde{\psi}_2(r) &= b_0 r + o(r) \end{aligned} \tag{5.47}$$

as $r \to 0^+$. The details of the proof are omitted. Now the initial conditions (5.43) and (5.44) imply that a_0 and b_0 satisfy

$$a_0 = p(0) + \lambda_1\frac{\nu}{\gamma} \quad \text{and} \quad b_0 = \lambda_2\frac{\nu}{\gamma}. \tag{5.48}$$

Subsituting (5.47) into (5.9), we obtain

$$\begin{aligned} \psi_1(r) &= r^3 + o(r^3), \\ \psi_2(r) &= O(r^3) \end{aligned} \tag{5.49}$$

as $r \to 0^+$.

The other properties of $\psi(\cdot,\lambda)$ can also be obtained as in proving Lemma 5.1; the details are omitted. The proof is complete. □

By arguing as in Theorems 5.4 and 5.6, we can obtain the following stability result for problem (EP).

THEOREM 5.8. *Let u be a neutrally stable B-type stationary solution of* (EP) *when $\nu = 0$. Then for any $\nu > 0$, u is stable with respect to* (5.43) *and* (5.15) *if $\kappa_1 \geq 0$ and $\kappa_2 \geq 0$. On the other hand, there is a positive constant $\kappa^* = \kappa^*(\nu, u)$ such that for any $\kappa_1 < -\kappa^*$, there is a nonempty open set $U(\kappa_1, \nu, u)$ such that u is unstable with respect to* (5.43) *and* (5.15) *for (κ_1, κ_2) with $\kappa_2 \in U(\kappa_1, \nu, u)$.*

Proof. The proof is the same as was used for Theorems 5.4 and 5.6. Therefore, the details are omitted. □

Remark 5.9. In their recent work on (EC), Makino et al. [15, 16, 19] showed that when $\gamma > \frac{4}{3}$ and $\nu > 0$, $u_{R,\mu}$ is nonlinearly asymptotically stable with respect to small perturbations. Their result is consistent with ours.

Appendix A. Asymptotic behavior at *R*. In this section, we shall study the asymptotic behavior of a real solution ψ at R for (3.3) with real ℓ. Let

$$\tau = R - r \quad \text{and} \quad \tilde{\psi}(\tau) = \psi(r).$$

Then $\tilde{\psi}$ satisfies

$$\tilde{\psi}'' + \{(1+q)\tau^{-1} + g(\tau)\}\tilde{\psi}' + \tau^{-1} f(\tau)\tilde{\psi} = 0.$$

For simplicity, we omit the $\sim$'s and write the last equation as

$$\text{(A.1)} \qquad \psi'' + \{(1+q)\tau^{-1} + g(\tau)\}\psi' + \tau^{-1} f(\tau)\psi = 0, \quad \tau > 0,$$

where g and f are continuous at $\tau = 0$.

Then we have the following result concerning the behavior of ψ at 0.

PROPOSITION A.1. *For any $q > 1$, let ψ be a solution of* (A.1). *Then either ψ is bounded at* 0 *or*

$$\text{(A.2)} \qquad \psi(\tau) = \tau^{-q}\hat{\psi}(\tau)$$

with $\hat{\psi}$ continuous at 0 *and $\hat{\psi}(0) \neq 0$. Furthermore, in the former case, ψ is C^2 at* 0*, and in the latter case, we have*

$$\text{(A.3)} \qquad \psi'(\tau) = -q\tau^{-q-1}\hat{\psi}(0) + o(\tau^{-q-1})$$

as $\tau \to 0^+$.

Proof. If g and f are analytic in a neighborhood of $\tau = 0$, then the result is well known; see, e.g., [2]. For completeness, we provide a proof here that assumes only that g and f are continuous at $\tau = 0$. Since the proof is elementary, some details are omitted.

For $\tau > 0$, let

$$\text{(A.4)} \qquad \psi(\tau) = \tau^{-q}\omega(\tau).$$

ω then satisfies

$$\text{(A.5)} \qquad \omega'' + \{(1-q)\tau^{-1} + g\}\omega' + (f - qg)\tau^{-1}\omega = 0, \quad \tau > 0.$$

Let $G(0) = 0$ and $G'(\tau) = g(\tau)$, (A.5) can then be written as

$$\text{(A.6)} \qquad (\tau^{1-q}e^G\omega')' + \tau^{-q}e^G(f - qg)\omega = 0.$$

Fix $\tau_1 > 0$ and let $\tau_0 \in (0, \tau_1)$ be chosen later. After integrating (A.6) from τ to τ_0, we have

$$\text{(A.7)} \qquad \omega'(\tau) = \tau^{q-1}E(\tau)C_0 + \tau^{q-1}E(\tau)\int_\tau^{\tau_0} F(s)s^{-q}\omega(s)ds,$$

where

$$E(\tau) = \exp(-G(\tau)), \qquad F(\tau) = (f - qg)\exp(G(\tau)),$$
$$\text{and} \quad C_0 = \tau_0^{1-q}\omega'(\tau_1)\exp(G(\tau_1)).$$

We first claim that

$$\omega \text{ and } \omega' \text{ are bounded on } (0,\tau]. \tag{A.8}$$

Indeed, let

$$C_1 = |C_0| \max_{\tau\in[0,\tau_1]} |E(\tau)| \quad \text{and} \quad C_2 = \max_{\tau\in[0,\tau_1]} |E(\tau)| \cdot \max_{\tau\in[0,\tau_1]} |F(\tau)|.$$

Then from (A.7), we have

$$|\omega'(\tau)| \le C_1\tau^{q-1} + C_2\tau^{q-1}\int_\tau^{\tau_0} s^{-q}|\omega(s)|ds, \tag{A.9}$$

which implies that

$$|\omega'(\tau)| \le C_1\tau^{q-1} + C_3 \max_{s\in[\tau,\tau_0]} |\omega(s)|,$$

where

$$C_3 = C_2\cdot(q-1)^{-1}.$$

Now for any $\tau \in (0,\tau_0)$, substituting (A.9) into

$$\omega(\tau) = \omega(\tau_0) + \int_{\tau_1}^{\tau} \omega'(s)ds,$$

we obtain

$$|\omega(\tau)| \le |\omega(\tau_0)| + C_4 + C_3\tau_0 \max_{s\in[\tau,\tau_0]} |\omega(s)|, \tag{A.10}$$

where

$$C_4 = C_1\tau_1^q.$$

Now if we choose $C_3\tau_0 < 1$, (A.10) then implies

$$|\omega(\tau)| \le (1 - C_3\tau_0)^{-1}\{|\omega(\tau_0)| + C_4\}.$$

Hence ω is bounded on $[0,\tau_1]$. By (A.9), ω' is bounded on $[0,\tau_1]$, which also implies that ω is continuous at 0. Now if $\omega(0) \ne 0$, then (A.4) and (A.8) imply (A.3). If $\omega(0) = 0$, we shall claim that

$$|\omega(\tau)| \le C_5\tau^q \tag{A.11}$$

for some $C_5 > 0$. Indeed, $\omega(0) = 0$ and ω' bounded on $[0,\tau_1]$ implies that

$$|\omega(\tau)| \le C_6\tau. \tag{A.12}$$

Now substituting (A.12) into (A.9), we have

$$|\omega'(\tau)| \le C_7\tau^{q-1} + C_8\tau \tag{A.13}$$

for some $C_7 > 0$ and $C_8 > 0$. Substituting (A.13) into

$$\omega(\tau) = \int_0^\tau \omega'(s)ds,$$

we obtain a better estimate for ω than (A.12). After repeating the processes a finite number of times, (A.15) follows. The proof is complete. □

Appendix B. Friedrichs' criteria. In this section, we recall a useful criterion of Friedrichs [4] for studying the spectra of second-order differential operators that are self-adjoint and singular at their endpoints.

Let $J = (x-, x+) \subset \mathbf{R}^1$ be a bounded or unbounded open interval.

Let $a(x)$, $a'(x)$, $b(x)$, and $c(x)$ be continuous functions on J. Furthermore, $a(x)$ and $c(x)$ are positive on J. The eigenvalue equation

$$-(a(x)\phi'(x))' + b(x)\phi(x) = \lambda c(x)\phi(x)$$

can be written as

$$L\phi = \lambda\phi,$$

where

$$L = c^{-1}(x)\left\{-\frac{d}{dx}\left(a(x)\frac{d}{dx}\right) + b(x)\right\}.$$

Define

$$h = \left|\int \frac{dx}{a(x)} + C\right| > 0$$

in the neighborhood of $x-$ or $x+$. The constants $C = C_-$ or $C = C_+$ should be chosen such that at the endpoint, h is either zero or infinite.

$$\text{(B.1)} \qquad \text{If } h(x-) = 0 \qquad (\text{or } h(x+) = 0),$$

then we require that

$$\text{(B.2)} \qquad \phi(x-) = 0 \qquad (\text{or } \phi(x+) = 0).$$

Otherwise, we need not put conditions on ϕ at $x-$ or $x+$.

Define $X = \{\phi : J \to \mathbf{R}^1 : \phi$ is absolutely continuous and satisfies $\int_{x-}^{x+} c(x)\phi^2(x)dx < \infty$, $\int_{x-}^{x+} a(x)\phi'^2(x)dx < \infty$, and $\int_{x-}^{x+} |b(x)|\phi^2(x)dx < \infty$ and also satisfies (B.1) if (B.2) holds$\}$ and

$$(\phi, \psi) = \int_{x-}^{x+} c(x)\phi(x)\psi(x)dx.$$

Definition B.1. *The spectrum of L is called* discrete below λ_* *if for every $\lambda' < \lambda_*$ there exists at most a finite number of mutually orthogonal eigenfunctions $\phi_\lambda(x)$ associated with eigenvalue $\lambda \le \lambda'$ such that for every $\phi \in X$ such that*

$$(\phi, \phi_\lambda) = 0,$$

we have

$$(\phi, L\phi) \ge \lambda_*(\phi, \phi).$$

L is called totally discrete *if L possesses a pure point spectrum.*

Remark B.2. If the spectrum is discrete below every λ_*, then it is totally discrete.

Define

$$Z(x) = \frac{1}{c(x)}\left\{b(x) + \frac{1}{4a(x)h^2(x)}\right\}.$$

Friedrichs' criterion can then be stated as follows.

Proposition B.3.

(I) *L is totally discrete if*

$$Z(x) \to \infty \quad \text{as } x \to x- \text{ and } x \to x+.$$

(II) *L is discrete below λ_* if*

$$\liminf Z(x) \geq \lambda_* \quad \text{as } x \to x- \text{ and } x \to x+.$$

(III) *L is not discrete below λ^* if $Z(x)$ is bounded below and*

$$\limsup Z(x) < \lambda^*$$

as either $x \to x-$ or $x \to x+$.

(IV) *The spectrum of L is discrete below λ_*, unbounded below, if*

$$\liminf_{x \to x-} Z(x) \geq \lambda_*,$$

$$\lim_{x \to x+} Z(x) = -\infty,$$

and

$$\Omega = \int_{x_0}^{x+} \left\{ \frac{c(x)}{-a(x)Z(x)} \right\}^{\frac{1}{2}} dx < \infty,$$

where $x_0 < x+$ such that $Z(x) < 0$ in $(x_0, x+)$. A similar result holds if the roles of $x-$ and $x+$ are interchanged.

Acknowledgments. The author wishes to thank Professors T. P. Liu and T. Makino for several useful discussions about this work.

REFERENCES

[1] S. Chandrasekhar, *An Introduction to the Study of Stellar Structures*, University of Chicago Press, Chicago, 1939.

[2] E. A. Coddington and N. Levison, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vols. I and II, Interscience, New York, 1953 and 1962.

[4] K. O. Friedrichs, *Criteria for discrete spectra*, Comm. Pure Appl. Math., 3 (1950), pp. 439–449.

[5] W. C. Kuan and S. S. Lin, *Numbers of equilibria for the equation of self-gravitating isentropic gas surrounding a solid ball*, Japan J. Indust. Appl. Math., 13 (1996), pp. 311–331.

[6] T. Makino, *On a local existence theorem for the evolution equation of gaseous stars*, in Patterns and Waves: Qualitative Analysis of Nonlinear Differential Equations, T. Nishida, M. Mimura, and K. Fujii, eds., North–Holland, Amsterdam, 1986, pp. 459–479.

[7] T. Makino, *Blowing up solutions of the Euler–Poisson equation for the evolution of gaseous stars*, Transport Theory Statist. Phys., 21 (1992), pp. 615–624.

[8] T. Makino, *Mathematical aspects of the Euler–Poisson equation for the evolution of gaseous stars*, lecture notes, National Chiao-Tung University, Hsin-chu, Taiwan, 1993.

[9] T. Makino, K. Mizohata, and S. Ukai, *The global weak solutions of compressible Euler equation with spherical symmetry*, Japan J. Indust. Appl. Math., 9 (1992), pp. 431–449.

[10] T. Makino, K. Mizohata, and S. Ukai, *The global weak solutions of compressible Euler equation with spherical symmetry* II, Japan J. Indust. Appl. Math., 11 (1994), pp. 417–426.

[11] T. Makino and B. Perthame, *Sur les solutions à symétric sphérique de l'equation d'Euler–Poisson pour l'evolution d'etoiles gazeuses*, Japan J. Appl. Math., 7 (1990), pp. 165–170.

[12] T. Makino and S. Takeno, *Initial boundary value problem for the spherically symmetric motion of isentropic gas*, Japan J. Indust. Appl. Math., 11 (1994), pp. 173–183.

[13] T. Makino and S. Ukai, *Sur l'existence des solutions locales de l'equation d'Euler–Poisson pour l'evolution gazeuses*, J. Math. Kyoto Univ., 27 (1987), pp. 387–399.

[14] T. Makino, S. Ukai, and S. Kawashima, *Sur la solution à support compact de l'equation d'Euler compressible*, Japan J. Appl. Math., 3 (1986), pp. 249–257.

[15] Š. Matušů-Nečasová, M. Okada, and T. Makino, *Free boundary problem for the equation of spherically symmetric motion of viscous gas* II, Japan J. Indust. Appl. Math., 12 (1995).

[16] Š. Matušů-Nečasová, M. Okada, and T. Makino, *Free boundary problem for the equation of spherically symmetric motion of viscous gas* III, preprint.

[17] W.-M. Ni and R. Nussbaum, *Uniqueness and non-uniqueness for positive radial solution of* $\Delta u + f(u, r) = 0$, Comm. Pure Appl. Math., 38 (1985), pp. 67–108.

[18] W.-M. Ni and J. Serrin, *Existence and nonexistence theorems for ground states of quasilinear partial differential equations: The anomalous case*, Acad. Naz. Lincei, 77 (1986), pp. 231–257.

[19] M. Okada and T. Makino, *Free boundary problem for the equation of spherically symmetric motion of viscous gas*, Japan J. Indust. Appl. Math., 10 (1993), pp. 219–235.

[20] A. D. Rendall, *The initial value problem for self-gravitating fluid bodies*, in Mathematical Physics X, K. Schmudgen, ed., Springer-Verlag, Berlin, 1992, pp. 471–474.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 570–594, May 1997

LONG-TIME BEHAVIOR FOR A CONVECTION-DIFFUSION EQUATION IN HIGHER DIMENSIONS[*]

MIGUEL ESCOBEDO[†] AND ENRIQUE ZUAZUA[‡]

Abstract. We study the long-time behavior of the solutions of the Cauchy problem

$$u_t - \triangle u + \frac{\partial |u|^{q-1}u}{\partial x_1} + \frac{\partial |u|^{p-1}u}{\partial x_2} = 0$$

in $\mathbf{R}^N \times (0,\infty)$ with initial data in $L^1(\mathbf{R}^N)$. We consider the range of exponents

$$1 < q < 1 + \frac{1}{N}, \qquad 1 + \frac{q}{N+1} < p.$$

We prove that nonnegative solutions with smooth initial data satisfy the entropy inequality

$$\frac{\partial u^{q-1}}{\partial x_1} \le \frac{C}{t}.$$

With the aid of this inequality, we show that, as $t \to \infty$, solutions behave like the nonnegative fundamental entropy solution of the reduced equation

$$u_t - \sum_{j=2}^{N} \frac{\partial^2 u}{\partial {x_j}^2} + \frac{\partial |u|^{q-1}u}{\partial x_1} = 0$$

of mass $\int_{\mathbf{R}^N} u(x,0)dx$, which has a self-similar structure.

Key words. parabolic scalar conservation law, convection-diffusion equation, entropy inequality, entropy solution, conservation of mass, long-time behavior, self-similarity

AMS subject classifications. 35K55, 35L65, 35B40, 35A08

PII. S0036141094271120

Introduction. In this paper, we continue the study of the long-time behavior of the solutions to some simple examples of scalar parabolic conservation laws, such as

$$u_t - \triangle u + \sum_{i=1}^{N} \frac{\partial f_i(u)}{\partial x_i} = 0 \quad \text{in } Q = \mathbf{R}^N \times (0,\infty), \tag{0.1}$$

$$u(\cdot,0) = u_0(\cdot) \quad \text{in } \mathbf{R}^N, \tag{0.2}$$

where $N \ge 1$, $f_i \in C^1([0,\infty)) \cap C^2((0,\infty))$, $f(0) = 0$, and $u_0 \in L^1(\mathbf{R}^N)$.

Solutions of (0.1)–(0.2) with L^1 initial data satisfy the following two properties:

$$\int_{\mathbf{R}^N} u(t,x)dx = \int_{\mathbf{R}^N} u_0(x)dx \quad \forall t \ge 0 \quad \text{(conservation of mass);} \tag{0.3}$$

[*] Received by the editors July 13, 1994; accepted for publication (in revised form) October 4, 1995. This research was supported by DGICYT (Spain) grant PB93-1203.
http://www.siam.org/journals/sima/28-3/27112.html

[†] Departamento de Matemáticas, Universidad del País Vasco, Apartado 644, E-48080 Bilbao, Spain (mtpesmam@lg.ehu.es). The research of this author was partially supported by Universidad del País Vasco (Spain) grant UPV 127.310-EA 134/93.

[‡] Departamento de Matemática Aplicada, Universidad Complutense, E-28040 Madrid, Spain (zuazua@sunma4.mat.ucm.es). The research of this author was partially supported by the Eurhomogenization project SC1*-CT91-0732.

$$\|u(t)\|_\infty \leq C\left(\int_{\mathbf{R}^N} u_0\right) t^{-\frac{N}{2}} \quad \forall t > 0 \quad (L^\infty \text{ decay}). \tag{0.4}$$

The asymptotic behavior of the solutions of (0.1)–(0.2) was studied in [EZ], [EVZ1], [EVZ2], [EVZ3], [C1], [C2], and [Z1] under one of the following conditions:

$$\text{for every } i \in \{1,\dots,N\}, \quad \text{the limits } \lim_{s\to 0} \frac{f_i(s)}{|s|^{1/N}s} \text{ exist} \tag{0.5}$$

or

$$\begin{cases} \exists q \in \left(1, 1+\dfrac{1}{N}\right) \quad \text{and} \quad j \in \{1,\dots,N\} \quad \text{such that } f_i \equiv 0 \quad \text{if } i \neq j, \\ f_j(0) = f_j'(0) = 0 \quad \text{and} \quad \text{the following limit exists:} \\ \lim\limits_{s\to 0} \dfrac{f_j''(s)}{|s|^{q-3}s}. \end{cases} \tag{0.6}$$

In these papers, three different types of asymptotic behaviors were observed: weakly nonlinear, self-similar, and strongly nonlinear. Let us briefly recall the main results obtained in these previous works.

Weakly nonlinear behavior. This case was considered in [EZ, section 6, Theorem 5], where the following result was proved. (See [Z2] for the second term of the asymptotic development.)

THEOREM 0.1. *Suppose that $f_i(s)/|s|^{1/N}s \to 0$ as $s \to 0$ for all $i = 1,\dots,N$ and assume that u solves* (0.1)–(0.2) *with $\int u_0 = M$. Then for every $r \in [1,\infty]$,*

$$\lim_{t\to\infty} t^{\frac{N}{2}(1-\frac{1}{r})} ||u(t) - MK(t)||_{L^r(\mathbf{R}^N)} = 0,$$

where $K(t,x)$ is the heat kernel in $\mathbf{R}^N$.

Self-similar behavior. It was proved in [AEZ] that if $q = 1 + 1/N$, then for every $M \in \mathbf{R}$ and every $\mathbf{C} \equiv (C_1,\dots,C_N) \in \mathbf{R}^N$, there is a unique self-similar solution $\omega_{M,\mathbf{C}}$ of the problem

$$\begin{cases} \omega_t - \triangle\omega + \sum\limits_{i=1}^{N} C_i \dfrac{\partial|\omega|^{q-1}\omega}{\partial x_i} = 0 \quad \text{for } t > 0, \quad x \in \mathbf{R}^N, \\ \omega(t,x) = t^{-N/2} g\left(\dfrac{x}{\sqrt{t}}\right); \quad \displaystyle\int g = M. \end{cases} \tag{0.7}$$

We refer to [AEZ] and [K] for further properties of the profiles g. In [EZ] (see also [Z1]), we then proved the following.

THEOREM 0.2. *Suppose that $\mathbf{f} \equiv (f_1,\dots,f_N)$ satisfies $f_i(s)/|s|^{1/N}s \to C_i$ as $s \to 0$ for all $i = 1,\dots,N$ with $\mathbf{C} = (C_1,\dots,C_N) \neq 0$, and assume that u is the solution of* (0.1)–(0.2) *with $\int_{\mathbf{R}^N} u_0 = M$. Then for every $r \in [1,\infty]$,*

$$\lim_{t\to\infty} t^{\frac{N}{2}(1-\frac{1}{r})} ||u(t) - \omega_{M,\mathbf{C}}(t)||_{L^r(\mathbf{R}^N)} = 0.$$

Strongly nonlinear behavior. It was proven in [C1] and [C2] that for every $M \in \mathbf{R}$ and $C \in \mathbf{R}$, there is a unique function $v_M \in BC((0,\infty); L^1(\mathbf{R}^N))$ which satisfies the "reduced equation"

$$v_t - \sum_{j=2}^{N} \frac{\partial^2 v}{\partial {x_j}^2} + C\frac{\partial|v|^{q-1}v}{\partial x_1} = 0 \quad \text{in } \mathcal{D}'((0,\infty)\times\mathbf{R}^N), \tag{0.8}$$

the initial conditions

$$
\begin{cases} \lim\limits_{t\to 0} \displaystyle\int_{\mathbf{R}^N} v(t,x)\phi(x)dx = M\phi(0) \quad \forall \phi \in BC(\mathbf{R}^N), \\ \forall r > 0, \quad \lim\limits_{t\to 0} \displaystyle\int_{\mathbf{R}^{N-1}} \int_{|x_1|>r} |u(t,x_1,\overline{x})| dx_1 d\overline{x} = 0 \end{cases} \tag{0.9}
$$

with $\overline{x} = (x_2, \ldots, x_N)$, and a suitable "entropy condition" (see Proposition 1.2 below for details).

By BC, we denote the set of bounded and continuous functions.

Note that equation (0.8) has a parabolic nature in the directions $\overline{x} = (x_2, \ldots, x_N)$ but a purely hyperbolic one in the direction x_1.

In [EVZ2], this uniqueness result was proved in the class of constant-sign solutions. Then the restriction on the sign was removed by Carpio [C1], [C2].

The entropy solution v_M of (0.8)–(0.9) has a self-similar structure and decays in $L^\infty(\mathbf{R}^N)$ like $t^{-(N+1)/2q}$, which is a faster decay rate than (0.4) since $1 < q < (N+1)/N$.

The following result about the asymptotic behavior of the solutions was obtained in [EVZ2] for constant-sign solutions and then extended to general solutions in [C1] and [C2].

Theorem 0.3. *Suppose that* $\mathbf{f} \equiv (f_1, \ldots, f_N)$ *satisfies* (0.6) *with* $j = 1$ *and*

$$
\lim_{s\to 0} \frac{f_1(s)}{|s|^{q-1}s} = C.
$$

Assume that u *is the solution of* (0.1)–(0.2) *with* $\int_{\mathbf{R}^N} u_0 = M$. *Then for every* $r \in [1,\infty)$,

$$
\lim_{t\to\infty} t^{\frac{N+1}{2q}(1-\frac{1}{r})} ||u(t) - v_M(t)||_{L^r(\mathbf{R}^N)} = 0,
$$

where v_M *is the fundamental entropy solution of* (0.8)–(0.9).

If we consider the simple example $f_i \equiv C_i|u|^{q_i-1}u$, condition (0.5) means that for every i, $q_i \geq 1 + 1/N$, while in the range $q \in (1, 1+1/N)$, (0.6) means that the convection is pointing in a fixed space direction.

In this paper, we consider a simple example where, in the frame of strongly nonlinear behavior, the convection is not unidirectional, namely

$$
u_t - \triangle u + \frac{\partial |u|^{q-1}u}{\partial x_1} + \frac{\partial |u|^{p-1}u}{\partial x_2} = 0 \tag{0.10}
$$

with

$$
1 < q < 1 + \frac{1}{N}, \quad q < p. \tag{0.11}
$$

Observe that the first-order, nonlinear term of (0.10) can be written as $q|u|^{q-1}\langle(1, (p/q)|u|^{p-q}, 0, \ldots, 0), \nabla u\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbf{R}^N$. It can therefore be seen as a convection term in the nonconstant direction $(1, (p/q)|u|^{p-q}, 0, \ldots, 0)$. Since the solutions of equation (0.10) with initial data $u_0 \in L^1(\mathbf{R}^N)$ decay in L^∞ (see (0.4)), as t goes to ∞, the vector that determines the direction of convection points increasingly in the direction $(1, 0, \ldots, 0)$. One is then tempted to conclude that the solutions u of (0.10) behave like the solutions of the equation

$$
u_t - \triangle u + \frac{\partial |u|^{q-1}u}{\partial x_1} = 0 \tag{0.12}
$$

and therefore, in view of Theorem 0.3, like the fundamental entropy solution of the reduced equation

$$u_t - \sum_{j=2}^{N} \frac{\partial^2 u}{\partial {x_j}^2} + \frac{\partial |u|^{q-1} u}{\partial x_1} = 0. \tag{0.13}$$

The main result of this paper shows that this is actually true under a further restriction on p.

Main result. *Suppose that*

$$1 < q < 1 + \frac{1}{N} \quad \textit{and} \quad 1 + \frac{q}{N+1} < p. \tag{0.14}$$

Assume that u is the solution of (0.10) *with initial data u_0 in $L^1(\mathbf{R}^N)$ of mass $\int_{\mathbf{R}^N} u_0 = M$. Then for every $r \in [1, \infty)$,*

$$\lim_{t \to \infty} t^{\frac{N+1}{2q}(1-\frac{1}{r})} ||u(t) - v_M(t)||_{L^r(\mathbf{R}^N)} = 0, \tag{0.15}$$

where v_M is the fundamental entropy solution of (0.13) *of mass M.*

Note that $1 + q/(N+1) > q$ if and only if $q < 1 + 1/N$. Also notice that no sign conditions are imposed to u_0.

Assumption (0.14) is probably sharp, i.e., the long-time behavior is probably of a different nature in the range $1 < q < 1 + 1/N$ and $q < p \leq 1 + q/(N+1)$. Although we do not have a rigorous proof of this fact, it will be obvious from the scaling arguments below. (See section 5.1 for a more detailed discussion.)

Notice that an immediate corollary of our main result is that solutions of (0.10) decay like $t^{-(N+1)/2q}$ in L^∞. Obtaining this decay rate is actually the hardest part of the proof of the main result. Once the decay estimate is obtained, the proof of convergence follows from a classical scaling argument where the compactness is derived by means of the kinetic approach developped by Lions, Perthame, and Tadmor [LPT] that was adapted to the present frame by Carpio [C1], [C2].

To get such a decay estimate, we will prove an entropy inequality of the following form:

$$\frac{\partial u^{q-1}}{\partial x_1} \leq \frac{C}{t},$$

which is rather classical in the context of scalar hyperbolic conservation laws and which was proved in [EVZ1] and [EVZ2] for parabolic conservation laws of the form (0.12).

The remainder of the paper is organized as follows.

Section 1 covers previous results. Section 2 deals with the entropy estimate and decay in L^∞. Section 3 contains the proof of the main result for positive solutions, and section 4 contains the proof of the main result for general solutions. Finally, further comments are presented in section 5.

1. Previous results. Throughout this paper, we will use the following result concerning the existence, uniqueness, and a priori estimates for classical solutions to (0.11) with initial data in $L^1(\mathbf{R}^N)$. Its proof is by now very classical. For a very similar result and its detailed proof, see, for instance, [EZ].

Proposition 1.1. *For all $p > 1$ and $q > 1$ and all initial data $u_0 \in L^1(\mathbf{R}^N)$, there is a unique, classical solution $u \in \mathbf{C}((0, \infty); L^1(\mathbf{R}^N))$ of* (0.11) *and* (0.2) *such that*

$$u \in \mathbf{C}((0, \infty); W^{2,r}(\mathbf{R}^N)) \cap \mathbf{C}^1((0, \infty); L^r(\mathbf{R}^N))$$

for every $r \in (1, +\infty)$. Moreover, this solution satisfies

(1.1)
$$\begin{cases} ||u(t)||_{L^1(\mathbf{R}^N)} \leq ||u_0||_{L^1(\mathbf{R}^N)}, \\ \forall r \in [1, +\infty], \quad \exists C_r \equiv C(r, ||u_0||_{L^1(\mathbf{R}^N)}) \quad ||u(t)||_{L^r(\mathbf{R}^N)} \leq C_r t^{-\frac{N}{2}(1-\frac{1}{r})} \quad \forall t > 0, \end{cases}$$

(1.2)
$$\forall u_0 \in L^1(\mathbf{R}^N) \cap L^r(\mathbf{R}^N), \quad ||u(t)||_{L^r(\mathbf{R}^N)} \leq \left(C_r t + ||u_0||_{L^r(\mathbf{R}^N)}^{-\frac{2r}{N(r-1)}} \right)^{-\frac{N}{2}(1-\frac{1}{r})} \qquad \forall t > 0,$$

$$\forall u_0 \in L^1(\mathbf{R}^N) \cap L^\infty(\mathbf{R}^N), \quad ||u(t)||_{L^\infty(\mathbf{R}^N)} \leq ||u_0||_{L^\infty(\mathbf{R}^N)} \quad \forall t > 0. \tag{1.3}$$

Here we also state the results that we need regarding the entropy solutions of the reduced equation (0.13).

In what follows, we will use the following notation: every $x \in \mathbf{R}^N$ is written as $x = (x_1, \overline{x})$, where $\overline{x} = (x_2, \ldots, x_N) \in \mathbf{R}^{N-1}$.

Proposition 1.2 (see [EVZ2], [C1], and [C2]). *If $1 < q < 1 + 1/N$, for every $M \in \mathbf{R}$, there is a unique function $v_M \in C((0, \infty); L^1(\mathbf{R}^N))$ that satisfies*

$$v_t - \sum_{j=2}^{N} \frac{\partial^2 v}{\partial {x_j}^2} + \frac{\partial |v|^{q-1} v}{\partial x_1} = 0 \qquad \text{in } \mathcal{D}'((0, \infty) \times \mathbf{R}^N), \tag{1.4}$$

$$\begin{cases} |v - \psi|_t - \displaystyle\sum_{j=2}^{N} \frac{\partial^2 |v - \psi|}{\partial {x_j}^2} + \frac{\partial ||v|^{q-1} v - |\psi|^{q-1} \psi|}{\partial x_1} \leq \operatorname{sign}(v - \psi) \sum_{j=2}^{N} \frac{\partial^2 \psi}{\partial {x_j}^2} \\ \quad \text{in } \mathcal{D}'((0, \infty) \times \mathbf{R}^N), \quad \forall \psi = \psi(\overline{x}) \in \mathcal{D}(\mathbf{R}^{N-1}), \end{cases} \tag{1.5}$$

$$\lim_{t \to 0} \int v(t, x) \phi(x) dx = M \phi(0) \quad \forall \phi \in BC(\mathbf{R}^N), \tag{1.6}$$

$$\forall r > 0, \quad \lim_{t \to 0} \int_{\mathbf{R}^{N-1}} \int_{|x_1| > r} |v(t, x_1, \overline{x})| dx_1 d\overline{x} = 0. \tag{1.7}$$

The function v_M has the self-similar form

$$v_M(t, x) = t^{-\frac{N+1}{2q}} g_M \left(\frac{x_1}{t^\beta}, \frac{\overline{x}}{\sqrt{t}} \right), \tag{1.8}$$

where

$$\beta = \frac{N + 1 + q - Nq}{2q} \tag{1.9}$$

and $g_M \in L^1(\mathbf{R}^N) \cap L^\infty(\mathbf{R}^N)$ has compact support in the x_1 variable and decays exponentially as $|x| \to \infty$.

Inequality (1.5) is an entropy-type condition in the spirit of Kruzhkov [Kr], which deals with the fact that equation (1.4) has lost the diffusion term in the x_1 variable. A weak solution of (1.4) that also satisfies (1.6) will be called an entropy solution.

2. Entropy estimate and decay in $L^\infty(\mathbf{R}^N)$. Observe that when $1 < q < 1 + 1/N$, $q < 1 + q/(N+1) < 1 + 1/N$ and therefore hypothesis (0.14) can be written as

$$1 < q < 1 + \frac{q}{N+1} < p. \tag{2.1}$$

After the results in [EVZ2], [C1], and [C2], it is clear that in this range of parameters, problem (0.10) does not have a parabolic scaling in all the variables. In order to make this statement more precise, let us proceed as follows. If u is the solution of (0.10) given by Proposition 1.1, we define for every $\lambda > 0$

$$u_\lambda(t, x_1, \overline{x}) = \lambda^{\frac{N+1}{2q}} u(\lambda t, \lambda^\beta x_1, \sqrt{\lambda}\overline{x}), \tag{2.2}$$

where

$$\beta = \frac{N + 1 + q - Nq}{2q}. \tag{2.3}$$

The function u_λ satifies

$$\frac{\partial u_\lambda}{\partial t} - \sum_{j=2}^{N} \frac{\partial^2 u_\lambda}{\partial x_j^2} + \frac{\partial |u_\lambda|^{q-1} u_\lambda}{\partial x_1} = \lambda^{1-2\beta} \frac{\partial^2 u_\lambda}{\partial x_1^2} - \lambda^{\frac{1}{2} - \frac{(N+1)(p-1)}{2q}} \frac{\partial |u_\lambda|^{p-1} u_\lambda}{\partial x_2}, \tag{2.4}$$

$$u_\lambda(0, x) = \lambda^{\frac{N+1}{2q}} u_0(\lambda^\beta x_1, \sqrt{\lambda}\overline{x}) \quad \forall x \equiv (x_1, \overline{x}) \in \mathbf{R}^N. \tag{2.5}$$

Observe that

$$\begin{cases} \beta > \dfrac{1}{2} \Longleftrightarrow q < 1 + \dfrac{1}{N}, \\ \dfrac{1}{2} - \dfrac{N+1}{2q}(p-1) < 0 \Longleftrightarrow 1 + \dfrac{q}{N+1} < p. \end{cases} \tag{2.6}$$

The idea of the proof of the main theorem is to pass to the limit in (2.4) as $\lambda \to \infty$, using (2.6), to say that in the limit we obtain a solution U of the problem (0.13) that can be identified with v_M. The main difficulty is to obtain the appropriate a priori bounds on the family $\{u_\lambda\}$. Observe that if we want $\{u_\lambda(t)\}$ to be uniformly bounded in $L^r(\mathbf{R}^N)$ for some $r \geq 1$ and $t > 0$, we need for u the estimate

$$||u(t)||_{L^r(\mathbf{R}^N)} \leq C t^{-\frac{N+1}{2q}(1-\frac{1}{r})}. \tag{2.7}$$

One of the main tools for doing this is the entropy estimate. The following theorem, which is the main result of this section, provides those two estimates.

THEOREM 2.1. *Suppose that p and q satisfy* (2.1). *Assume that u is the solution of* (0.10) *and* (0.2) *with nonnegative initial data u_0 of mass $\int_{\mathbf{R}^N} u_0 dx = M$ such that*

$$u_0 \in L^1(\mathbf{R}^N) \cap L^\infty(\mathbf{R}^N). \tag{2.8}$$

Then there is a positive constant $C \equiv C(p, q, ||u_0||_{L^\infty(\mathbf{R}^N)}, M)$ such that

$$\left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \leq C t^{-1} \quad \forall t > 0 \tag{2.9}$$

and

$$||u(t)||_{L^\infty(\mathbf{R}^N)} \le Ct^{-\frac{N+1}{2q}} \quad \forall t > 0. \tag{2.10}$$

Theorem 2.1 will be proved via a sequence of lemmas.

LEMMA 2.2. *Assume that the hypotheses of Theorem* 2.1 *hold. Then for some positive t_0 and some positive constant $A \equiv A(p,q)$, we have*

$$\left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \le At^{-1} \quad \forall t \in (0, t_0) \quad \forall x \in \mathbf{R}^N. \tag{2.11}$$

Moreover, if for some positive constants C_0 and α, we have

$$u(t,x) \le C_0(1+t)^{-\alpha} \quad \forall t > 0 \quad \forall x \in \mathbf{R}^N, \tag{2.12}$$

then for every $\tau_0 > 0$, there is a positive constant $C = C(\tau_0, p, q, C_0, \alpha)$ such that

$$\left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \le Ct^{-\sigma} \quad \forall t > \tau_0 \quad \forall x \in \mathbf{R}^N \tag{2.13}$$

with $\sigma = \min(1, 2\alpha(p-1))$.

Proof. By the comparison principle, since $u_0 \ge 0$, the solution u is strictly positive in $(0,\infty) \times \mathbf{R}^N$. Define $z = qu^{q-1}$. The equation (0.10) then reads

$$z_t - \triangle z - \gamma \frac{|\nabla z|^2}{z} + z\frac{\partial z}{\partial x_1} + q(q-1)u^{q-2}\frac{\partial u^p}{\partial x_2} = 0$$

with $\gamma = (2-q)/(q-1) > 0$. If we differentiate in x_1 and denote $w = \partial z/\partial x_1$, we have

$$\begin{aligned} &w_t - \triangle w + w^2 + \gamma\frac{|\nabla z|^2}{z^2}w + \left(z - 2\gamma\frac{w}{z}\right)\frac{\partial w}{\partial x_1} - 2\frac{\gamma}{z}\sum_{j=2}^N \frac{\partial z}{\partial x_j}\frac{\partial w}{\partial x_j} \\ &\quad + q(q-1)\frac{\partial}{\partial x_1}\left(u^{q-2}\frac{\partial u^p}{\partial x_2}\right) = 0. \end{aligned} \tag{2.14}$$

Now the last term in the left-hand side is

$$q(q-1)\frac{\partial}{\partial x_1}\left(u^{q-2}\frac{\partial u^p}{\partial x_2}\right) = C_1(p,q)z^{\frac{p-1}{q-1}}\frac{\partial w}{\partial x_2} + C_2(p,q)\frac{z^{\frac{p-1}{q-1}}}{z}w\frac{\partial z}{\partial x_2}.$$

Therefore, w satisfies

$$\begin{aligned} &w_t - \triangle w + w^2 + \gamma\frac{|\nabla z|^2}{z^2}w + \left(z - 2\gamma\frac{w}{z}\right)\frac{\partial w}{\partial x_1} \\ &\quad - 2\frac{\gamma}{z}\sum_{j=2}^N \frac{\partial z}{\partial x_j}\frac{\partial w}{\partial x_j} + C_1(p,q)z^{\frac{p-1}{q-1}}\frac{\partial w}{\partial x_2} + C_2(p,q)\frac{z^{\frac{p-1}{q-1}}}{z}w\frac{\partial z}{\partial x_2} = 0. \end{aligned} \tag{2.15}$$

This is a parabolic equation whose coefficients are not bounded in all of $(0,\infty) \times \mathbf{R}^N$ since $u(t,x)$ tends to zero as $|x| \to \infty$ at any positive time t. Nevertheless, since $u > 0$ and u is smooth for $t > 0$, $w \in L^\infty_{\text{loc}}((0,\infty) \times \mathbf{R}^N) \cap \mathbf{C}^{2,1}((0,\infty) \times \mathbf{R}^N)$.

Let us prove that for $t > 0$, $w(t) \in L^\infty(\mathbf{R}^N)$ and satisfies the estimate (2.11). To this end, consider the solution u_ε of (0.10) with initial data $u_\varepsilon(0) = u_0 + \varepsilon$ and $0 < \varepsilon < 1$. By the uniqueness of solutions and the classical parabolic regularity, we deduce that for every $t > 0$, $u_\varepsilon(t) \to u(t)$ and $\nabla u_\varepsilon(t) \to \nabla u(t)$ uniformly on any compact subset of $\mathbf{R}^N$ as $\varepsilon \to 0$.

On the other hand, since $u_0 \geq 0$, $u_\varepsilon \geq \varepsilon$. Therefore, the function $w_\varepsilon \equiv q(u_\varepsilon^{q-1})_{x_1}$ is in $L^\infty((0,T) \times \mathbf{R}^N)$ for every $\varepsilon > 0$ and $T > 0$. Moreover, it satisfies (2.15) with $z_\varepsilon = qu_\varepsilon^{q-1}$ instead of z. Therefore, multiplying the equation satisfied by w_ε by $\operatorname{sign}^+ w_\varepsilon$, using Kato's inequality, and taking into account that

$$\begin{aligned}\left| \frac{z_\varepsilon^{\frac{p-1}{q-1}}}{z_\varepsilon} w_\varepsilon^+ \frac{\partial z_\varepsilon}{\partial x_2} \right| &\leq w_\varepsilon^+ \left\{ \delta \left| \frac{\partial z_\varepsilon}{\partial x_2} \right|^2 z_\varepsilon^{-2} + \frac{1}{4\delta} z_\varepsilon^{\frac{2(p-1)}{q-1}} \right\} \\ &\leq \delta w_\varepsilon^+ \left| \frac{\partial z_\varepsilon}{\partial x_2} \right|^2 z_\varepsilon^{-2} + \delta |w_\varepsilon^+|^2 + \frac{1}{64\delta^3} z_\varepsilon^{\frac{4(p-1)}{q-1}}\end{aligned}$$

for any $\delta > 0$, by choosing $\delta > 0$ small enough, we obtain
(2.16)

$$\begin{aligned}\frac{\partial w_\varepsilon^+}{\partial t} - \triangle w_\varepsilon^+ + \frac{1}{2}(w_\varepsilon^+)^2 + \frac{\gamma}{2}\frac{|\nabla z_\varepsilon|^2}{z_\varepsilon^2} w_\varepsilon^+ + \left(z_\varepsilon - 2\gamma \frac{w_\varepsilon}{z_\varepsilon} \right) \frac{\partial w_\varepsilon^+}{\partial x_1} - 2\frac{\gamma}{z_\varepsilon} \sum_{j=2}^N \frac{\partial z_\varepsilon}{\partial x_j} \frac{\partial w_\varepsilon^+}{\partial x_j} \\ + C_1(p,q) z_\varepsilon^{\frac{p-1}{q-1}} \frac{\partial w_\varepsilon^+}{\partial x_2} \leq C(p,q) ||u_\varepsilon(t)||_{L^\infty(\mathbf{R}^N)}^{4(p-1)} \leq C(p,q)(||u_0||_{L^\infty(\mathbf{R}^N)} + 1)^{4(p-1)}\end{aligned}$$

for all $0 < \varepsilon < 1$ since $||u_\varepsilon(t)||_{L^\infty(\mathbf{R}^N)} \leq ||u_\varepsilon(0)||_{L^\infty(\mathbf{R}^N)} = ||u_0||_{L^\infty(\mathbf{R}^N)} + \varepsilon$.

Let us set $t_0 = (1/2)(||u_0||_{L^\infty(\mathbf{R}^N)} + 1)^{-2(p-1)}$ and define $W_\varepsilon(\tau; t) = A(t + \tau)^{-1}$ for every $\varepsilon > 0$, where $\tau \equiv \tau(\varepsilon)$ and $A > 0$ is independent of ε such that

(i) $A^2/2 - A \geq C(p,q)$ (with $C(p,q)$ as in the right-hand side of (2.16)),

(ii) $0 < \tau < t_0$, and

(iii) $A/\tau > ||w_\varepsilon(0)||_{L^\infty(\mathbf{R}^N)}$.

A simple calculation shows that

$$\frac{dW_\varepsilon}{dt} + \frac{1}{2}W_\varepsilon^2 = \left(\frac{A^2}{2} - A \right)(\tau + t)^{-2} \geq C(p,q)(\tau + t)^{-2} \quad \forall t > 0.$$

Now if $t \in (0, t_0)$, we have

$$(\tau + t) \leq 2t_0 = (||u_0||_{L^\infty(\mathbf{R}^N)} + 1)^{-2(p-1)}.$$

Then

$$(\tau + t)^{-2} \geq (||u_0||_{L^\infty(\mathbf{R}^N)} + 1)^{4(p-1)},$$

and so

$$\frac{dW_\varepsilon}{dt} + \frac{1}{2}W_\varepsilon^2 = \left(\frac{A^2}{2} - A \right)(\tau + t)^{-2} \geq 2C(p,q)(||u_0||_{L^\infty(\mathbf{R}^N)} + 1)^{4(p-1)} \quad \forall t \in (0, t_0).$$

Since $W_\varepsilon(0) > w_\varepsilon(0)$ by construction, by the comparison principle, we deduce that

$$w_\varepsilon^+(t,x) \leq A(t + \tau(\varepsilon))^{-1} \leq At^{-1} \quad \forall x \in \mathbf{R}^N \quad \forall t \in (0, t_0).$$

Passing to the limit as $\varepsilon \to 0$, we obtain (2.11).

In order to prove (2.13), we use the following argument. Since equation (0.10) is autonomous, for every $\tau \in (0, t_0/2)$, the function defined by $\widetilde{u}(t,x) = u(t + \tau, x)$ solves (0.10) with initial data $\widetilde{u}(0) = u(\tau)$ and satisfies (2.12). Therefore, the function $\widetilde{w} = q\partial\widetilde{u}^{q-1}/\partial x_1$ solves an equation similar to (2.15), which for the sake of brevity we call $\widetilde{(2.15)}$. In the same way, we can define $\widetilde{u}_\varepsilon$, $\widetilde{z}_\varepsilon$, and $\widetilde{w}_\varepsilon$ by $\widetilde{z}_\varepsilon(t) = z_\varepsilon(t+\tau)$ and $\widetilde{w}_\varepsilon(t) = w_\varepsilon(t+\tau)$, where $\widetilde{u}_\varepsilon$ is the solution of (0.10) with data $u(\tau)+\varepsilon$. Then

$$\begin{aligned} \frac{\partial \widetilde{w}_\varepsilon^+}{\partial t} - \triangle \widetilde{w}_\varepsilon^+ + \frac{1}{2}(\widetilde{w}_\varepsilon^+)^2 + \frac{\gamma}{2}\frac{|\nabla \widetilde{z}_\varepsilon|^2}{\widetilde{z}_\varepsilon^2}(\widetilde{w}_\varepsilon^+) + \left(\widetilde{z}_\varepsilon - 2\gamma\frac{\widetilde{w}_\varepsilon}{\widetilde{z}_\varepsilon}\right)\frac{\partial \widetilde{w}_\varepsilon^+}{\partial x_1}& \\ - 2\frac{\gamma}{\widetilde{z}_\varepsilon}\sum_{j=2}^{N} \frac{\partial \widetilde{z}_\varepsilon}{\partial x_j}\frac{\partial \widetilde{w}_\varepsilon^+}{\partial x_j} + C_1(p,q)\widetilde{z}_\varepsilon^{\frac{p-1}{q-1}}\frac{\partial \widetilde{w}_\varepsilon^+}{\partial x_2} &\le C(p,q)(||u_0||_{L^\infty(\mathbf{R}^N)}+1)^{4(p-1)}. \end{aligned}$$

Since $\widetilde{w_\varepsilon}(0) = w_\varepsilon(\tau)$ by (2.11), $\widetilde{w}_\varepsilon(0) \in L^\infty(\mathbf{R}^N)$. On the other hand, by what we have just proved, $||\widetilde{w}_\varepsilon(0)||_{L^\infty(\mathbf{R}^N)} < A\tau^{-1}$ for every $\varepsilon > 0$. By the comparison principle, we deduce that

$$\forall \varepsilon > 0, \quad \forall t > 0, \quad \forall x \in \mathbf{R}^N, \quad \widetilde{w}_\varepsilon(t,x) \le C(p,q)(||u_0||_{L^\infty(\mathbf{R}^N)}+1)^{4(p-1)}t + A\tau^{-1},$$

and so, letting $\varepsilon \to 0$, $\widetilde{w} \in L^\infty_{\text{loc}}((0,\infty); L^\infty(\mathbf{R}^N))$.

Moreover, multiplying $\widetilde{(2.15)}$ by $\text{sign}^+\widetilde{w}$ and using (2.12), we get

$$\left\{ \begin{aligned} \frac{\partial \widetilde{w}^+}{\partial t} - \triangle \widetilde{w}^+ + \frac{1}{2}(\widetilde{w}^+)^2 + \frac{\gamma}{2}\frac{|\nabla \widetilde{z}|^2}{(\widetilde{z})^2}(\widetilde{w}^+) + (\widetilde{z} - 2\gamma\frac{\widetilde{w}}{\widetilde{z}})\frac{\partial \widetilde{w}^+}{\partial x_1}& \\ - 2\frac{\gamma}{\widetilde{z}}\sum_{j=2}^{N} \frac{\partial \widetilde{z}}{\partial x_j}\frac{\partial \widetilde{w}^+}{\partial x_j} + C_1(p,q)\widetilde{z}^{\frac{p-1}{q-1}}\frac{\partial \widetilde{w}^+}{\partial x_2} &\le C(p,q)(1+t)^{-4\alpha(p-1)}. \end{aligned} \right.$$

We now consider the cases where $2\alpha(p-1) \ge 1$ and $2\alpha(p-1) < 1$ separately.

The case where $2\alpha(p-1) \ge 1$. Consider the function $W(t) = A(t+\tau_1)^{-1}$ with A such that $A^2/2 - A > C(p,q)$ and $1 > \tau_1 > 0$ such that $||\widetilde{w}(0)||_{L^\infty(\mathbf{R}^N)} < A/\tau_1$. With this choice of A and τ_1, we have

$$\frac{dW}{dt} + \frac{1}{2}W^2 = \left(\frac{A^2}{2} - A\right)(\tau_1 + t)^{-2} > C(p,q)(1+t)^{-4\alpha(p-1)}.$$

By the comparison principle, we deduce as before that

$$w^+(t+\tau, x) \equiv \widetilde{w}^+(t,x) \le A(t+\tau_1)^{-1} < At^{-1} \quad \forall t > 0 \quad \forall x \in \mathbf{R}^N.$$

The case where $2\alpha(p-1) < 1$. Define the function $W(t) = A(1+t)^{-2\alpha(p-1)}$, choosing A such that

$$A\left(\frac{A}{2} - 2\alpha(p-1)\right) \ge 2C(p,q) \quad \text{and} \quad A \equiv A(\tau) > ||\widetilde{w}(0)||_{L^\infty(\mathbf{R}^N)}.$$

Then

$$\begin{aligned} \frac{dW}{dt} + \frac{1}{2}W^2 &= A(1+t)^{-4\alpha(p-1)}\left\{\frac{A}{2} - 2\alpha(p-1)(1+t)^{2\alpha(p-1)-1}\right\} \\ &\ge A(1+t)^{-4\alpha(p-1)}\left\{\frac{A}{2} - 2\alpha(p-1)\right\} \ge 2C(p,q)(1+t)^{-4\alpha(p-1)}. \end{aligned}$$

Arguing as before, we obtain $\widetilde{w}^+(t) \le A(1+t)^{-2\alpha(p-1)}$ for all $t > 0$. This gives

$$\left[\frac{\partial u^{q-1}(t+\tau, x)}{\partial x_1}\right]^+ \le A(1+t)^{-\sigma} \quad \forall t > 0$$

and, for some $C = C(\sigma, \tau_0) > 0$, $[\partial u^{q-1}/\partial x_1]^+ \leq Ct^{-\sigma}$ for every $t > \tau_0$, and (2.13) follows. □

LEMMA 2.3. *Assume that the hypotheses of Theorem* 2.1 *are satisfied and that* (2.12) *holds. Then we have the following:*

(a) *If* $2\alpha(p-1) < 1$, *for every* $\delta > 0$, *there exists* $C \equiv C(p, q, \delta, C_0) > 0$ *such that*

$$\left\| \int_{\mathbf{R}} u(t, z, \overline{x}) dz \right\|_{L^\infty(\mathbf{R}^{N-1})} \leq Ct^{-\alpha(N-1)(p-1)+\delta} \quad \forall t > 2^{N-1}. \tag{2.17}$$

(b) *If* $2\alpha(p-1) \geq 1$, *there exists* $C \equiv C(p, q, C_0) > 0$ *such that*

$$\left\| \int_{\mathbf{R}} u(t, z, \overline{x}) dz \right\|_{L^\infty(\mathbf{R}^{N-1})} \leq Ct^{-\frac{N-1}{2}} \quad \forall t > 2^{N-1}. \tag{2.18}$$

Proof. Let us define $v(t, \overline{x}) = \int_{\mathbf{R}} u(t, z, \overline{x}) dz$. This function satisfies

$$v_t - \sum_{j=2}^{N} \frac{\partial^2 v}{\partial {x_j}^2} + \frac{\partial \int_{\mathbf{R}} u^p(t, z, \overline{x}) dz}{\partial x_2} = 0,$$

and therefore it solves the integral equation

$$v(2t, \overline{x}) = (K_{N-1}(t) * v(t))(\overline{x}) - \int_0^t \left(\frac{\partial}{\partial x_2} K_{N-1}(t-s) * \int_{\mathbf{R}} u^p(s+t, z, \cdot) dz \right) (\overline{x}) ds, \tag{2.19}$$

where K_{N-1} denotes the heat kernel in $(0, \infty) \times \mathbf{R}^{N-1}$, i.e.,

$$K_{N-1}(t, \overline{x}) = (4\pi t)^{-\frac{N-1}{2}} \exp\left(-\frac{|\overline{x}|^2}{4t} \right).$$

The following estimates are easily obtained:

$$||K_{N-1}(t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma_0 t^{-\frac{N-1}{2}(1-\frac{1}{r})},$$
$$||\nabla K_{N-1}(t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma_0 t^{-\frac{N-1}{2}(1-\frac{1}{r})-\frac{1}{2}}$$

for all $r \in [1, \infty]$ and $t > 0$ with $\gamma_0 > 0$ large enough.

Now consider any r such that

$$1 < r < \frac{N-1}{N-2} \quad \text{if } N > 2, \qquad r > 1 \quad \text{if } N = 2$$

and take $L^r(\mathbf{R}^{N-1})$ norms in (2.19). Using (2.12), we obtain

$$\begin{aligned} ||v(2t)||_{L^r(\mathbf{R}^{N-1})} &\leq \gamma_0 M t^{-\frac{N-1}{2}(1-\frac{1}{r})} \\ &\quad + \gamma_0 M C_0^{p-1} \int_0^t (t-s)^{-\frac{N-1}{2}(1-\frac{1}{r})-\frac{1}{2}} (1+s+t)^{-\alpha(p-1)} ds, \end{aligned}$$

and then

$$||v(2t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma_0 M t^{-\frac{N-1}{2}(1-\frac{1}{r})} + \frac{2r\gamma_0 M C_0^{p-1}}{r-(N-1)(r-1)} t^{-\frac{N-1}{2}(1-\frac{1}{r})+\frac{1}{2}-\alpha(p-1)}.$$

If $2\alpha(p-1) \geq 1$, $1/2 - \alpha(p-1) \leq 0$ and then

$$||v(t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r})} \quad \forall t \geq 2 \tag{2.20}$$

with $\gamma_1 = M\gamma_0(1 + 2rC_0^{p-1}/(r - (N-1)(r-1)))$.

On the other hand, if $2\alpha(p-1) < 1$,

$$||v(t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r})+\frac{1}{2}-\alpha(p-1)} \quad \forall t \geq 2. \tag{2.21}$$

When $N = 2$, this gives an estimate on the $L^r(\mathbf{R}^N)$ norm of $v(t)$ for every finite $r > 1$. When $N > 2$, we divide the interval $(1/(N-1), 1)$ into the intervals $(j/(N-1), (j+1)/(N-1))$, $j = 1, \ldots, N-2$. Now for any $r > (N-1)$, choose $(r_j)_{j=1}^{N-2}$ such that $1/r_j \in ((N-j-1)/(N-1), (N-j)/(N-1))$ and

$$\frac{1}{r_{N-2}} - \frac{1}{r} < \frac{1}{N-1}, \tag{2.22}$$

$$\frac{1}{r_j} - \frac{1}{r_{j+1}} < \frac{1}{N-1} \quad \forall j = 1, \ldots, N-2. \tag{2.23}$$

We can now use (2.20)–(2.21) with $r_1 \in (1, (N-1)/(N-2))$ and take the $L^{r_2}(\mathbf{R}^{N-1})$ norm in (2.19):

$$\begin{aligned} ||v(2t)||_{L^{r_2}(\mathbf{R}^{N-1})} &\leq M\gamma_0 t^{-\frac{N-1}{2}(\frac{1}{r_1}-\frac{1}{r_2})} ||v(t)||_{L^{r_1}(\mathbf{R}^{N-1})} \\ &\quad + \left\| \int_0^t \left(\frac{\partial}{\partial x_2} K_{N-1}(t-s) * \int_{\mathbf{R}} u^p(s+t, z, \cdot) dz \right) ds \right\|_{L^{r_2}(\mathbf{R}^{N-1})}. \end{aligned}$$

The last term in the right-hand side can be estimated as follows for $1/r_1 + 1/\overline{r}_1 = 1 + 1/r_2$:

$$\begin{aligned} &\left\| \int_0^t \left(\frac{\partial}{\partial x_2} K_{N-1}(t-s) * \int_{\mathbf{R}} u^p(s+t, z, \cdot) dz \right) (\overline{x}) ds \right\|_{L^{r_2}(\mathbf{R}^{N-1})} \\ &\quad \leq \int_0^t ||u(s+t)||_{L^\infty(\mathbf{R}^N)}^{p-1} \left\| \left| \frac{\partial}{\partial x_2} K_{N-1}(t-s) \right| * v(s+t) \right\|_{L^{r_2}(\mathbf{R}^{N-1})} ds \\ &\quad \leq \int_0^t ||u(s+t)||_{L^\infty(\mathbf{R}^N)}^{p-1} \left\| \frac{\partial}{\partial x_2} K_{N-1}(t-s) \right\|_{L^{\overline{r}_1}(\mathbf{R}^{N-1})} ||v(s+t)||_{L^{r_1}(\mathbf{R}^{N-1})} ds. \end{aligned}$$

First, suppose that $2\alpha(p-1) \geq 1$. Then using (2.12) and (2.20), we obtain, for all $t \geq 2$,

$$\begin{aligned} ||v(2t)||_{L^{r_2}(\mathbf{R}^{N-1})} &\leq \gamma_0 t^{-\frac{N-1}{2}(\frac{1}{r_1}-\frac{1}{r_2})} \gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r_1})} \\ &\qquad + \gamma_0 C_0^{p-1} \gamma_1 \int_0^t (t-s)^{-\frac{N-1}{2}(\frac{1}{r_1}-\frac{1}{r_2})-\frac{1}{2}} (s+t)^{-\frac{N-1}{2}(1-\frac{1}{r_1})-\alpha(p-1)} ds \\ &= \gamma_0\gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r_2})} \\ &\qquad + \gamma_0 C_0^{p-1} \gamma_1 \frac{2r_1 r_2}{r_1 r_2 - (N-1)(r_2 - r_1)} t^{-\frac{N-1}{2}(1-\frac{1}{r_2})+(\frac{1}{2}-\alpha(p-1))}. \end{aligned}$$

Since $2\alpha(p-1) \geq 1$, we deduce that

$$||v(t)||_{L^{r_2}(\mathbf{R}^{N-1})} \leq \gamma_2 t^{-\frac{N-1}{2}(1-\frac{1}{r_2})} \quad \forall t \geq 4 \tag{2.24}$$

with $\gamma_2 = \gamma_0\gamma_1(1 + C_0^{p-1} 2r_1 r_2/(r_1 r_2 - (N-1)(r_2 - r_1)))$.

On the other hand, suppose that $2\alpha(p-1) < 1$. Then using (2.20), we obtain, for all $t \geq 2$,

$$\begin{aligned}
||v(2t)||_{L^{r_2}(\mathbf{R}^{N-1})} &\leq \gamma_0\gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r_2})+\frac{1}{2}-\alpha(p-1)} \\
&\quad + \gamma_0 C_0^{p-1}\gamma_1 \int_0^t (t-s)^{-\frac{N-1}{2}(\frac{1}{r_1}-\frac{1}{r_2})-\frac{1}{2}}(s+t)^{-\frac{N-1}{2}(1-\frac{1}{r_1})+\frac{1}{2}-2\alpha(p-1)}ds \\
&= \gamma_0\gamma_1 t^{-\frac{N-1}{2}(1-\frac{1}{r_2})+\frac{1}{2}-\alpha(p-1)} \\
&\quad + \gamma_0 C_0^{p-1}\gamma_1 \frac{2r_1r_2}{r_1r_2-(N-1)(r_2-r_1)} t^{-\frac{N-1}{2}(1-\frac{1}{r_2})+2(\frac{1}{2}-\alpha(p-1))}.
\end{aligned}$$

Since $2\alpha(p-1) < 1$, we deduce that

$$(2.25) \qquad ||v(t)||_{L^{r_2}(\mathbf{R}^{N-1})} \leq \gamma_2 t^{-\frac{N-1}{2}(1-\frac{1}{r_2})+2(\frac{1}{2}-\alpha(p-1))} \quad \forall t \geq 4.$$

By iteration of this argument, we obtain

$$(2.26) \qquad ||v(t)||_{L^{r_j}(\mathbf{R}^{N-1})} \leq \gamma_j t^{-\theta_j} \quad \forall t \geq 2^j$$

for $j = 1, \ldots, N-2$, where

$$\theta_j \equiv \begin{cases} \dfrac{N-1}{2}\left(1-\dfrac{1}{r_j}\right) - j\left(\dfrac{1}{2}-\alpha(p-1)\right) & \text{if } 2\alpha(p-1) < 1, \\ \dfrac{N-1}{2}\left(1-\dfrac{1}{r_j}\right) & \text{if } 2\alpha(p-1) \geq 1 \end{cases}$$

and $\gamma_j = \gamma_0\gamma_1\cdots\gamma_{j-1}(1 + C_0^{p-1}2r_{j-1}r_j/(r_{j-1}r_j - (N-1)(r_j - r_{j-1})))$. Finally, we take L^r norms in (2.19) and use (2.20)–(2.21) for $j = N-2$ and (2.22) to get

$$(2.27) \qquad ||v(t)||_{L^r(\mathbf{R}^{N-1})} \leq \gamma t^{-\theta} \quad \forall t \geq 2^{N-1},$$

where

$$\theta \equiv \begin{cases} \dfrac{N-1}{2}\left(1-\dfrac{1}{r}\right) - (N-1)\left(\dfrac{1}{2}-\alpha(p-1)\right) & \text{if } 2\alpha(p-1) < 1, \\ \dfrac{N-1}{2}\left(1-\dfrac{1}{r}\right) & \text{if } 2\alpha(p-1) \geq 1 \end{cases}$$

and $\gamma = \gamma_0\gamma_1\cdots\gamma_{N-2}(1 + C_0^{p-1}2r_{N-2}r/(r_{N-2}r - (N-1)(r - r_{N-2})))$.

The estimate in L^∞ remains to be proved. Unfortunately, we cannot let $r \to \infty$ in (2.27) since (due to (2.22)–(2.23)) when that happens, $r_j \to (N-1)/(N-j-1)$ for all $j = 1, \ldots, N-2$ and then all of the constants γ_j blow up.

We first consider the case where $2\alpha(p-1) \geq 1$. For any $r > N-1$, taking L^∞ norms in (2.19), we obtain

$$\begin{aligned}
&||v(2t)||_{L^\infty(\mathbf{R}^{N-1})} \\
&\quad \leq \gamma_0 t^{-\frac{N-1}{2r}}\gamma t^{-\frac{N-1}{2}(1-\frac{1}{r})} + \gamma_0 C_0^{p-1}\gamma \int_0^t (t-s)^{-\frac{N-1}{2r}-\frac{1}{2}}(s+t)^{-\frac{N-1}{2}(1-\frac{1}{r})-\alpha(p-1)}ds \\
&\quad \leq \gamma_0\gamma\left(1 + \frac{2rC_0^{p-1}}{r-(N-1)}\right)t^{-(N-1)/2}
\end{aligned}$$

for every $t \geq 2^{N-1}$. This concludes the proof of Lemma 2.3 when $2\alpha(p-1) \geq 1$.

Let us now consider the case where $2\alpha(p-1) < 1$. We will use the following inequality of Gagliardo and Nirenberg (see, for instance, [B, p. 195]):

$$(2.28) \qquad \forall \rho \geq 1, \quad \forall r > (N-1), \quad ||v||_{L^\infty(\mathbf{R}^{N-1})} \leq C||v||^{1-a}_{L^\rho(\mathbf{R}^{N-1})}||v||^a_{W^{1,r}(\mathbf{R}^{N-1})},$$

where

$$(2.29) \qquad a\left(\frac{1}{\rho} - \frac{1}{r} + \frac{1}{(N-1)}\right) = \frac{1}{\rho}$$

and the constant C depends only on ρ and r. We will apply (2.28) to $v(t)$ with ρ large. We then need to estimate $||v(t)||_{W^{1,r}}$ for some $r > (N-1)$. Since we have already estimated $||v(t)||_{L^r(\mathbf{R}^{N-1})}$ in (2.27), we only have to estimate $||\nabla v(t)||_{L^r(\mathbf{R}^{N-1})}$. Since by definition $v(t, \overline{x}) = \int_{\mathbf{R}} u(t, z, \overline{x})dz$, we have $\nabla v(t, \overline{x}) = \int_{\mathbf{R}} \nabla_{\overline{x}} u(t, z, \overline{x})dz$. Let us define the auxiliary function $V(t, \overline{x}) = \int_{\mathbf{R}} |\nabla_{\overline{x}} u(t, z, \overline{x})|dz$. Then, obviously,

$$\forall t > 0, \quad \forall \overline{x} \in \mathbf{R}^{N-1}, \quad |\nabla v(t, \overline{x})| \leq V(t, \overline{x}).$$

To estimate $||V(t)||_{L^r(\mathbf{R}^{N-1})}$, we note that for all $t > 0$ and $\tau > 0$,

$$u(t+\tau) = K(t) * u(\tau) - \int_0^t \frac{\partial}{\partial x_1} K(t-s) * u^q(s+\tau)ds - \int_0^t \frac{\partial}{\partial x_2} K(t-s) * u^p(s+\tau)ds.$$

Taking the gradient in the $\overline{x}$ variables and norms in $\mathbf{R}^{N-1}$ and integrating in $\mathbf{R}$ with respect to x_1, we obtain

$$\begin{aligned} V(t+\tau, \overline{x}) \leq & \int_{\mathbf{R}} (|\nabla_{\overline{x}} K(t)| * u(\tau))dx_1 \\ & + q \int_{\mathbf{R}} \int_0^t \left|\frac{\partial}{\partial x_1} K(t-s)\right| * u^{q-1} |\nabla_{\overline{x}} u(s+\tau)| ds dx_1 \\ & + p \int_{\mathbf{R}} \int_0^t \left|\frac{\partial}{\partial x_2} K(t-s)\right| * u^{p-1} |\nabla_{\overline{x}} u(s+\tau)| ds dx_1. \end{aligned}$$

If we denote by $K_1(t, x_i)$ the heat kernel in the one space variable x_i, i.e., $K_1(t, x_i) = (4\pi t)^{-1/2} \exp(-|x_i|^2/4t)$, then $K(t, x) = K_1(t, x_1) \cdots K_1(t, x_N)$ and therefore $|\nabla_{\overline{x}} K(t, x)| = K_1(t, x_1)|\nabla K_{N-1}(t, \overline{x})|$. It is then straightforward to see that $\int_{\mathbf{R}} |\nabla_{\overline{x}} K(t)| * u(\tau)dx_1 = |\nabla K_{N-1}(t)| * v(\tau)$. On the other hand,

$$\begin{aligned} & \int_{\mathbf{R}} \left|\frac{\partial}{\partial x_1} K(t-s)\right| * u^{q-1} |\nabla_{\overline{x}} u(s+\tau)| dx_1 \\ & \qquad = \int_{\mathbf{R}^{N-1}} K_{N-1}(t-s, \overline{x} - \overline{y}) \\ & \int_{\mathbf{R}} \int_{\mathbf{R}} \frac{|x_1 - y_1|}{2(t-s)} K_1(t-s, x_1 - y_1) u^{q-1}(s+\tau, y) |\nabla_{\overline{x}} u(s+\tau, y)| dy_1 dx_1 d\overline{y} \\ & \qquad \leq \left(\int_{\mathbf{R}} \frac{|z|}{2(t-s)} K_1(t-s, z)dz\right) C_0^{q-1} (s+\tau)^{-\alpha(q-1)} \\ & \int_{\mathbf{R}^{N-1}} K_{N-1}(t, \overline{x} - \overline{y}) \int_{\mathbf{R}} |\nabla_{\overline{x}} u(s+\tau, y)| dy_1 d\overline{y} \\ & \qquad \leq \frac{C(q, \tau)}{\sqrt{t-s}} K_{N-1}(t-s) * V(s+\tau) \end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathbf{R}} \left|\frac{\partial}{\partial x_2} K(t-s)\right| * u^{p-1}|\nabla_{\overline{x}} u(s+\tau)|dx_1 \\
&\qquad = \int_{\mathbf{R}^{N-1}} \left|\frac{\partial}{\partial x_2} K_{N-1}(t-s, \overline{x}-\overline{y})\right| \\
&\int_{\mathbf{R}}\int_{\mathbf{R}} K_1(t-s, x_1-y_1) u^{p-1}(s+\tau, y)|\nabla_{\overline{x}} u(s+\tau, y)|dy_1 dx_1 d\overline{y} \\
&\qquad \leq C_0^{p-1}(s+\tau)^{-\alpha(p-1)} \int_{\mathbf{R}^{N-1}} \left|\frac{\partial}{\partial x_2} K_{N-1}(t-s, x-\overline{y})\right| \int_{\mathbf{R}} |\nabla_{\overline{x}} u(s+\tau, y)|dy_1 d\overline{y} \\
&\qquad \leq C(p,\tau) \left|\frac{\partial}{\partial x_2} K_{N-1}(t-s)\right| * V(s+\tau).
\end{aligned}$$

For all $t > 0$ and $\tau > 0$, we deduce that

$$||V(t+\tau)||_{L^r(\mathbf{R}^{N-1})} \leq Ct^{-\frac{1}{2}}||v(\tau)||_{L^r(\mathbf{R}^{N-1})} + C(p,q,\tau)\int_0^t (t-s)^{-\frac{1}{2}}||V(s+\tau)||_{L^r(\mathbf{R}^{N-1})}.$$

For any $\tau > 0$ fixed, if we set $g(t) = ||V(t+\tau)||_{L^r(\mathbf{R}^{N-1})}$, we can write, for every $t \in (0,1)$,

$$g(t) \leq t^{-\frac{1}{2}}||v(\tau)||_{L^r(\mathbf{R}^{N-1})} + C\int_0^t (t-s)^{-\frac{1}{2}} g(s)ds.$$

By Gronwall's lemma, we deduce that there is a positive constant $C' = C'(p,q,C_0)$ such that $g(t) \leq C'||v(\tau)||_{L^r(\mathbf{R}^{N-1})}t^{-1/2}$ for every $t \in (0,1)$. In particular, for every $\tau > 0$,

$$||\nabla v(1+\tau)||_{L^r(\mathbf{R}^{N-1})} \leq ||V(1+\tau)||_{L^r(\mathbf{R}^{N-1})} \equiv g(1) \leq C'||v(\tau)||_{L^r(\mathbf{R}^{N-1})}.$$

By (2.27), there is a positive constant Γ such that for all $t \geq 2^N$,

$$||v(t)||_{W^{1,r}(\mathbf{R}^{N-1})} \leq \Gamma t^{-\theta}. \tag{2.30}$$

We use now (2.27), (2.28), and (2.30) to obtain

$$||v(t)||_{L^\infty(\mathbf{R}^{N-1})} \leq C(\rho, r)\gamma^{1-a}\Gamma^a t^{-\Theta} \quad \forall t \geq 2^N \tag{2.31}$$

with

$$\begin{aligned}
\Theta \equiv (1-a)&\left[\frac{N-1}{2}\left(1-\frac{1}{\rho}\right) - (N-1)\left(\frac{1}{2}-\alpha(p-1)\right)\right] \\
&+ a\left[\frac{N-1}{2}\left(1-\frac{1}{r}\right) - (N-1)\left(\frac{1}{2}-\alpha(p-1)\right)\right],
\end{aligned}$$

which we can rewrite as

$$\begin{aligned}
\Theta \equiv (1-a)(N-1)\alpha(p-1) - (1-a)\frac{N-1}{2\rho} \\
+ a\left[\frac{N-1}{2}\left(1-\frac{1}{r}\right) - (N-1)\left(\frac{1}{2}-\alpha(p-1)\right)\right].
\end{aligned}$$

By (2.29), we have $a \equiv a(\rho, r) = (1 - \rho/r + \rho/(N-1))^{-1}$. Now we leave $r > N-1$ fixed and let $\rho \to \infty$. Then $a(\rho, r) \to 0$ as $\rho \to \infty$ and

$$\lim_{\rho\to\infty} \Theta = \begin{cases} (N-1)\alpha(p-1) & \text{if } 2\alpha(p-1) < 1, \\ \dfrac{N-1}{2} & \text{if } 2\alpha(p-1) \geq 1. \end{cases}$$

This completes the proof of Lemma 2.3 when $2\alpha(p-1) < 1$. □

LEMMA 2.4. *Assume that the hypotheses of Theorem* 2.1 *are satisfied and that for some positive constants l, τ_1, and C_1,*

$$\left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \leq C_1 t^{-l} \quad \forall t > \tau_1.$$

Then for every $\overline{x} \in \mathbf{R}^{N-1}$,

$$\sup_{x_1 \in \mathbf{R}} |u(t, x_1, \overline{x})| \leq \left(\frac{q}{q-1}\right)^{\frac{1}{q}} C_1^{\frac{1}{q}} t^{-\frac{l}{q}} \left(\int_{\mathbf{R}} u(t, z, \overline{x}) dz\right)^{\frac{1}{q}} \quad \forall t > \tau_1. \tag{2.32}$$

Proof. We follow the argument of [EVZ2]. We fix a time $t > \tau_1$ and consider a point $(y_1, \overline{y}, t)$. At this point, we call $B = u^{q-1}(y_1, \overline{y}, t)$. By hypothesis, we have

$$\forall x_1 \in \left[y_1 - \frac{B}{C_1}t^l, y_1\right], \quad u^{q-1}(x_1, \overline{y}, t) \geq B - C_1\frac{(y_1 - x_1)}{t^l}$$

or, equivalently,

$$u(y_1 - z, \overline{y}, t) \geq \left(\frac{C_1}{t^l}\right)^{\frac{1}{q-1}} \left(\frac{Bt^l}{C_1} - z\right)^{\frac{1}{q-1}} \quad \forall z \in \left[0, \frac{Bt^l}{C_1}\right].$$

Integrating with respect to z on $(0, Bt^l/C_1)$ gives

$$\begin{aligned} \int_{\mathbf{R}} u(z, \overline{y}, t) dz &\geq \left(\frac{C_1}{t^l}\right)^{\frac{1}{q-1}} \int_0^{\frac{Bt^l}{C_1}} \left(\frac{Bt^l}{C_1} - z\right)^{\frac{1}{q-1}} dz \\ &= \left(\frac{C_1}{t^l}\right)^{\frac{1}{q-1}} \int_0^{\frac{Bt^l}{C_1}} s^{\frac{1}{q-1}} ds = \frac{q-1}{q} B^{\frac{q}{q-1}} \frac{t^l}{C_1} \quad \forall t > \tau_1. \end{aligned}$$

Therefore,

$$u^q(y_1, \overline{y}, t) \leq \frac{q}{q-1} C_1 t^{-l} \int_{\mathbf{R}} u(z, \overline{y}, t) dz \quad \forall t > \tau_1. \qquad \square$$

Proof of Theorem 2.1. By (2.8), applying estimates (1.1) and (1.3) of Proposition 1.1, we have that for some positive constant C_0, the solution u satisfies

$$u(t, x) \leq C_0(1+t)^{-\frac{N}{2}} \quad \forall t > 0 \quad \forall x \in \mathbf{R}^N. \tag{2.33}$$

We can then apply Lemma 2.2 with $\alpha = N/2$. Thus for every $\tau_0 > 0$, there is a positive constant $C = C(\tau_0, p, q, C_0)$ such that

$$\left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \leq Ct^{-\sigma} \quad \forall t > \tau_0 \tag{2.34}$$

with $\sigma = \min(1, N(p-1))$.

From (2.34) and (2.11), we immediately see that if $N(p-1) \geq 1$, (2.9) is proved. Moreover, by Lemma 2.3 and, more precisely, by (2.18),

$$\int_{\mathbf{R}} u(z, \overline{y}, t)dz \leq Ct^{-\frac{N-1}{2}} \quad \forall t > 2^{N-1}$$

for some other constant $C = C(p, q, C_0)$. Then by Lemma 2.4,

$$\begin{aligned}|u(t, x_1, \overline{x})| &\leq \left(\frac{q}{q-1}\right)^{\frac{1}{q}} C^{\frac{1}{q}} t^{-\frac{1}{q}} \left(\int_{\mathbf{R}} u(t, z, \overline{x})dz\right)^{\frac{1}{q}} \\ &\leq \left(\frac{q}{q-1}\right)^{\frac{1}{q}} C^{\frac{1}{q}} t^{-\frac{1}{q}-\frac{N-1}{2q}} \quad \forall t > 2^N \quad \forall x \in \mathbf{R}^N.\end{aligned}$$

Taking into account the fact that (1.1) holds, this implies (2.10) since $(N+1)/2q > N/2$.

We can now suppose that $N(p-1) < 1$. Then $\sigma = N(p-1)$ and by Lemma 2.3, for every $\delta > 0$, there exists $C \equiv C(p, q, \delta, C_0) > 0$ such that

$$\left\|\int_{\mathbf{R}} u(t, z, \overline{x})dz\right\|_{L^\infty(\mathbf{R}^{N-1})} \leq Ct^{-\frac{N(N-1)(p-1)}{2}+\delta} \quad \forall t > 2^{N-1}. \tag{2.35}$$

We now choose δ in the following way. Since $p > 1 + q/(N+1)$, we have $(N+1)(p-1)/q > 1$. Therefore, for k large enough,

$$\frac{(p-1)^2(N+1)N}{q}\left(\frac{(N+1)(p-1)}{q}\right)^{k-1} > 2.$$

Let us set

$$k_0 = \min\left\{k \in \mathbf{N} : \frac{(p-1)^2(N+1)N}{q}\left(\frac{(N+1)(p-1)}{q}\right)^{k-1} > 2\right\}. \tag{2.36}$$

We choose δ such that

$$\begin{cases} \dfrac{(N+1)N(p-1)}{2q} - k_0\dfrac{\delta}{q} > \dfrac{(N+1)N(p-1)}{4q}, \\ \delta < \dfrac{N(N-1)(p-1) - Nq}{2}. \end{cases} \tag{2.37}$$

Now using (2.34), (2.35) (with δ defined by (2.37)), and Lemma 2.4,

$$\begin{cases} u(t, x) \leq Ct^{-\alpha_1} \quad \forall t > \tau \quad \forall x \in \mathbf{R}^N, \\ \alpha_1 = \dfrac{N(N+1)(p-1)}{2q} - \dfrac{\delta}{q} \end{cases} \tag{2.38}$$

for some $\tau > 0$. Using Lemma 2.2, this in turn gives

$$\forall \tau_0 > 0, \quad \exists C > 0 \quad \left[\frac{\partial u^{q-1}}{\partial x_1}\right]^+ \leq Ct^{-\sigma_1} \quad \forall t > \tau_0 \tag{2.39}$$

with $\tau > 0$ and $\sigma_1 = \min(1, 2\alpha_1(p-1))$. Therefore, if $2\alpha_1(p-1) \geq 1$, in view of (2.11) and (2.39), (2.9) is proved and (2.10) follows as above. On the other hand, if

$2\alpha_1(p-1) < 1$, then $\sigma_1 = 2\alpha_1(p-1)$, and then by (2.39) and (2.17), for every $\delta > 0$, there exists $C \equiv C(p, q, \delta, C_0) > 0$ such that

$$\left\| \int_{\mathbf{R}} u(t, z, \overline{x}) dz \right\|_{L^\infty(\mathbf{R}^{N-1})} \leq C t^{-\alpha_1 (N-1)(p-1)+\delta_1} \quad \forall t > 2^{N-1}. \tag{2.40}$$

We choose δ_1 to be

$$\delta_1 \equiv \delta \frac{(N+1)(p-1)}{q}. \tag{2.41}$$

Again by (2.39)–(2.41) and Lemma 2.4, we obtain

$$\begin{cases} u(t,x) \leq C t^{-\alpha_2} \quad \forall t > 0 \quad \forall x \in \mathbf{R}^N, \\ \alpha_2 = \dfrac{(N+1)(p-1)}{q} \left(\dfrac{N(N+1)(p-1)}{2q} - 2\dfrac{\delta}{q} \right). \end{cases}$$

In this way, step by step, by choosing $\delta_j = \delta_{j-1}(N+1)(p-1)/q$ we obtain a sequence of numbers

$$\alpha_k = \left[\frac{(N+1)(p-1)}{q} \right]^{k-1} \left(\frac{N(N+1)(p-1)}{2q} - k\frac{\delta}{q} \right)$$

such that if $2\alpha_{j-1}(p-1) < 1$, then

$$u(t,x) \leq C t^{-\alpha_j} \quad \forall t > 0 \quad \forall x \in \mathbf{R}^N,$$
$$\int_{\mathbf{R}} u(t, z, \overline{x}) dz \leq t^{-\alpha_{j-1}(N-1)(p-1)+\delta_{j-1}} \quad \forall t > 2^{N-1} \quad \forall x \in \mathbf{R}^N.$$

We claim that for $k = k_0$ (where k_0 is defined by (2.36)), we have $2\alpha_{k_0}(p-1) > 1$. For this observe that by the choice of δ (see (2.37)),

$$\begin{aligned} \alpha_{k_0} &= \left(\frac{N(N+1)(p-1)}{2q} - k_0 \frac{\delta}{q} \right) \left[\frac{(N+1)(p-1)}{q} \right]^{k_0 - 1} \\ &\geq \frac{N(N+1)(p-1)}{4q} \left(\frac{(N+1)(p-1)}{q} \right)^{k_0-1}, \end{aligned}$$

and so by the choice of k_0,

$$2\alpha_{k_0}(p-1) \geq \frac{(p-1)^2(N+1)N}{q} \left[\frac{(N+1)(p-1)}{q} \right]^{k_0-1} > 1.$$

Thus (2.9) and (2.10) are deduced as above. □

Corollary 2.5. *Under the hypotheses of Theorem* 2.1*, there is a positive constant C such that*

$$\frac{\partial u^q}{\partial x_1} \leq C \frac{u}{1+t} \quad \forall t > 0, \tag{2.42}$$

$$\int_{\mathbf{R}^N} \left| \frac{\partial u^q}{\partial x_1} \right| dx \leq \frac{CM}{(1+t)} \quad \forall t > 0. \tag{2.43}$$

Proof. Inequality (2.42) is a direct consequence of Theorem 2.1.

For (2.43), we use the fact that

$$\int_{\mathbf{R}^N} \frac{\partial u^q}{\partial x_1} dx = 0.$$

Then by (2.42),

$$\int_{\mathbf{R}^N} \left(\frac{\partial u^q}{\partial x_1}\right)^- dx = \int_{\mathbf{R}^N} \left(\frac{\partial u^q}{\partial x_1}\right)^+ \leq \frac{C}{1+t} \int_{\mathbf{R}^N} u dx. \qquad \square$$

3. Main result for nonnegative solutions. In this section, we prove the main result for nonnegative initial data. By the maximum principle, this means that the solution u is also nonnegative. In order to pass to the limit in (2.4) as $\lambda \to \infty$, we need some further estimates on the family $\{u_\lambda\}$ defined in (2.2).

LEMMA 3.1. *If u solves* (0.10) *and* (0.2) *and u_λ is the family of functions defined by* (2.2)*, then*

$$\begin{aligned} &\int_\tau^T \int_{\mathbf{R}^N} \left(|\nabla_{\overline{x}} u_\lambda(t,x)|^2 + \lambda^{1-2\beta} \left| \frac{\partial u_\lambda(s+\tau)}{\partial x_1} \right|^2 \right) dx dt \\ &\qquad \leq \frac{1}{2} \int_{\mathbf{R}^N} |u_\lambda(\tau,x)|^2 dx \leq C(M) \tau^{-\frac{(N+1)}{2q}} \quad \forall \lambda > 0 \end{aligned} \tag{3.1}$$

for all $T > 0$ and $\tau \in (0,T)$.

Proof. Multiplying equation (0.10) by u and integrating by parts, we get that for every $T > 0$ and $\tau \in (0,T)$,

$$\int_\tau^T \int_{\mathbf{R}^N} |\nabla u(t,x)|^2 dx dt \leq \frac{1}{2} \int_{\mathbf{R}^N} |u(\tau,x)|^2 dx \leq C(M) \tau^{-\frac{(N+1)}{2q}}.$$

The result follows from the definition of u_λ. $\square$

LEMMA 3.2. *If u solves* (0.10) *and* (0.2) *and u_λ is the family of functions defined by* (2.2)*, then for every $\tau > 0$, there is a positive constant such that*

$$||\nabla_{\overline{x}} u_\lambda(t)||_{L^1(\mathbf{R}^N)} \leq C \quad \forall t \geq \tau_0, \tag{3.2}$$

where $\nabla_{\overline{x}} u = (\partial u/\partial x_2, \dots, \partial u/\partial x_N)$.

Proof. For every λ, the function u_λ satisfies

$$\begin{aligned} u_\lambda(t+\tau) = K_\lambda(t) * u_\lambda(\tau) &- \int_0^t K_\lambda(t-s) * \left(\frac{\partial u_\lambda^q(s+\tau)}{\partial x_1} \right) ds \\ &- \lambda^{\frac{1}{2} - \frac{(N+1)(p-1)}{2q}} \int_0^t \left(\frac{\partial K_\lambda(t-s)}{\partial x_2} \right) * u_\lambda^p(s+\tau) ds, \end{aligned}$$

where $K_\lambda(t,x) = \lambda^{\beta-1/2} K(t, \lambda^{\beta-1/2} x_1, \overline{x})$. Then for every $t \in (0,1)$ and $\tau > \tau_0$,

$$\begin{aligned} &||\nabla_{\overline{x}} u_\lambda(t+\tau)||_{L^1(\mathbf{R}^N)} \\ &\quad \leq M ||\nabla_{\overline{x}} K_\lambda(t)||_{L^1(\mathbf{R}^N)} + \int_0^t ||\nabla_{\overline{x}} K_\lambda(t-s)||_{L^1(\mathbf{R}^N)} \left\| \frac{\partial u_\lambda^q(s+\tau)}{\partial x_1} \right\|_{L^1(\mathbf{R}^N)} ds \\ &\qquad + \lambda^{\frac{1}{2} - \frac{(N+1)(p-1)}{2q}} \int_0^t \left\| \frac{\partial K_\lambda(t-s)}{\partial x_2} \right\|_{L^1(\mathbf{R}^N)} ||\nabla_{\overline{x}} u_\lambda^p(s+\tau)||_{L^1(\mathbf{R}^N)} ds \\ &\quad \leq C t^{-\frac{1}{2}} + C \int_0^t (t-s)^{-\frac{1}{2}} (s+\tau)^{-1} ds \end{aligned}$$

$$+\int_0^t (t-s)^{-\frac{1}{2}}(s+\tau)^{-\frac{(N+1)(p-1)}{2q}}||\nabla_{\overline{x}} u_\lambda^p(s+\tau)||_{L^1(\mathbf{R}^N)}ds.$$

If $g(t)=||\nabla_{\overline{x}}u_\lambda(t+\tau)||_{L^1(\mathbf{R}^N)}$, we have for $t\in(0,1)$ and for all $\tau>\tau_0$ that

$$g(t)\le (C\sqrt{t}+CMt^{-\frac{1}{2}})+C\tau_0^{-\frac{(N+1)(p-1)}{2q}}\int_0^t (t-s)^{-\frac{1}{2}}g(s)ds.$$

From Gronwall's lemma, we deduce that (3.2) holds. □

LEMMA 3.3. *There is a sequence* $\{\lambda_n\}$ *and* $U\in L^\infty((0,\infty);L^1(\mathbf{R}^N))\cap L^\infty((0,\infty)\times\mathbf{R}^N)$*, a solution of* (1.4)*, satisfying the entropy condition* (1.5) *and such that*

$$\forall\varepsilon>0,\quad \forall t_2>t_1>0,\quad u_{\lambda_n}\to U\quad \text{in } \mathbf{C}([t_1,t_2];W^{-\varepsilon,2}_{\text{loc}}(\mathbf{R}^N)),$$
$$\forall r\in[1,\infty),\quad \forall t>0,\quad u_{\lambda_n}(t)\to U(t)\quad \text{in } L^r_{\text{loc}}(\mathbf{R}^N).$$

Proof. By Theorem 2.1 and Corollary 2.5, if u_0 is nonnegative and satisfies (2.8), the family u_λ defined by (2.2) satisfies

$$0<u_\lambda(t,x)\le Ct^{-\frac{N+1}{2q}}\quad \forall t>0\quad \forall x\in\mathbf{R}^N, \tag{3.3}$$

$$\frac{\partial u_\lambda^{q-1}(t)}{\partial x_1}\le Ct^{-1}\quad \text{in } (0,+\infty)\times\mathbf{R}^N, \tag{3.4}$$

$$\int_{\mathbf{R}^N}\left|\frac{\partial u_\lambda^q(t)}{\partial x_1}\right|dx\le CMt^{-1}\quad \forall t>0, \tag{3.5}$$

$$\int_{\mathbf{R}^N}\left|\frac{\partial u_\lambda^{q+r}(t)}{\partial x_1}\right|dx\le CMt^{-1-r\frac{N+1}{2q}}\quad \forall t>0. \tag{3.6}$$

The argument now follows [EVZ2]:

(i) From (3.2)–(3.3) and (3.6), we deduce that the family $\{u_\lambda^{q+r}\}$ is uniformly bounded in $L^\infty(\tau,\infty;W^{1,1}(\mathbf{R}^N))$ for every $r\ge 1$ and $\tau>0$.

(ii) From (3.1), (3.2), (3.5), and (2.4), we deduce that $\{\partial_t u_\lambda\}$ is uniformly bounded in $L^2_{\text{loc}}((0,\infty);H^{-s}_{\text{loc}}(\mathbf{R}^N))$ for some positive s.

(iii) From (i), we deduce that $\{u_\lambda\}$ is uniformly bounded in $L^\infty_{\text{loc}}((0,\infty);L^2_{\text{loc}}(\mathbf{R}^N))$.

For every bounded set Ω of $\mathbf{R}^N$, since $L^2(\Omega)$ is compactly embedded in $H^{-\varepsilon}(\Omega)$ for every $\varepsilon>0$ and $H^{-\varepsilon}(\Omega)$ is continuously embedded in $H^{-s}(\Omega)$ for every $s>\varepsilon$, we deduce from (ii), (iii), and the compactness result of [S] that

(iv) $\{u_\lambda^{q+r}\}$ is relatively compact in $\mathbf{C}([t_1,t_2];H^{-\varepsilon}(\Omega))$.

We can therefore extract a subsequence $\lambda_n\to\infty$ such that

$$u_{\lambda_n}\to U\quad \text{in } \mathbf{C}([t_1,t_2];H^{-\varepsilon}(\Omega)) \tag{3.7}$$

for every bounded domain Ω in $\mathbf{R}^N$ and every $\varepsilon<s$. Since by (i) and (3.3), we know that $\{u_\lambda(t)\}$ is relatively compact in $L^r_{\text{loc}}(\mathbf{R}^N)$ for every $1\le r<\infty$ and every $t>0$, we deduce that

$$u_{\lambda_n}(t)\to U(t)\quad \text{in } L^r_{\text{loc}}(\mathbf{R}^N). \tag{3.8}$$

Then passing to the limit in (2.4) through the sequence $\{\lambda_n\}$ in the sense of the distributions $\mathcal{D}'((0,\infty)\times\mathbf{R}^N)$, we see that U is a weak solution of (1.4). Moreover, since for every n, u_{λ_n} is a regular solution of (0.10), for any $\psi \in \mathcal{D}(\mathbf{R}^{N-1})$, $\psi = \psi(\overline{x})$, we have

$$
\begin{aligned}
&|u_{\lambda_n} - \psi|_t - \widetilde{\triangle}|u_{\lambda_n} - \psi| + \frac{\partial}{\partial x_1}|u^q_{\lambda_n} - |\psi|^{q-1}\psi| \\
&\quad \leq \operatorname{sign}(u_{\lambda_n} - \psi)\widetilde{\triangle}\psi + \lambda^{1-2\beta}\frac{\partial^2 u_{\lambda_n}}{\partial x_1^2}\operatorname{sign}(u_{\lambda_n} - \psi) - \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}}\frac{\partial|u^p_{\lambda_n} - |\psi|^{p-1}\psi|}{\partial x_2}
\end{aligned}
$$

in the sense of the distributions $\mathcal{D}'((0,\infty) \times \mathbf{R}^N)$, where $\tilde{\Delta}$ denotes the Laplacian in the variables $(x_2, \dots, x_N)$. Taking into account that

$$
\frac{\partial^2 u_{\lambda_n}}{\partial x_1^2}\operatorname{sign}(u_{\lambda_n} - \psi) \leq \frac{\partial^2 |u_{\lambda_n} - \psi|}{\partial x_1^2} + \frac{\partial^2 \psi}{\partial x_1^2}\operatorname{sign}(u_{\lambda_n} - \psi)
$$

and passing to the limit as $n \to \infty$, we obtain that U satisfies the entropy condition (1.5). Moreover, U satisfies (3.3) and (3.4). We then have that $U \in L^\infty((\tau,\infty) \times \mathbf{R}^N)$ for every $\tau > 0$, and since $||u(t)||_{L^1(\mathbf{R}^N)} = M$, we also have that $U \in L^\infty((0,\infty); L^1(\mathbf{R}^N))$ and $\int_{\mathbf{R}^N} U(t,x)dx \leq M$ $\forall t > 0$.

In order to identify U as the function v_M of Proposition 1.2, we also need to prove that U takes the initial data $M\delta$ in the sense of (1.6)–(1.7). To this end, we need the following.

Lemma 3.4. *Suppose that u is the classical solution of* (0.10) *and* (0.2) *and the family* $\{u_\lambda\}$ *is the one defined in* (2.2). (*Note that we do not assume u to be of constant sign.*) *Then for every $t_0 > 0$ and $\varepsilon > 0$, there are $k_0 > 0$ and $\lambda_0 > 0$ such that*

$$
\int_{|x_1|+|\overline{x}|\geq k_0} |u_\lambda|(t, x_1, \overline{x})dx_1 d\overline{x} \leq \varepsilon \tag{3.9}
$$

for every $t \in (0, t_0)$ and $\lambda \geq \lambda_0$.

Proof. If $N \geq 3$, we define

$$
v_\lambda(t, x_3, \dots, x_N) = \int_{\mathbf{R}^2} u_\lambda(t, x_1, x_2, x_3, \dots, x_N)dx_1 dx_2.
$$

For every $\lambda > 0$, the function v_λ satisties the $(N-2)$-dimensional linear heat equation. Since the family of initial data $\{v_\lambda(0)\}$ is an approximation of the identity in $\mathbf{R}^{N-2}$, we deduce that for every $t_2 > t_1 > 0$ and every $\varepsilon > 0$, there are $k_0 > 0$ and $\lambda_0 > 0$ such that

$$
\begin{aligned}
&\int_{|(x_3,\dots,x_N)|>k_0} v_\lambda(t, x_3, \dots, x_N)dx_3 \cdots dx_N \\
&\qquad \equiv \int_{\mathbf{R}^2}\int_{|(x_3,\dots,x_N)|>k_0} u_\lambda(t, x_3, \dots, x_N)dx_1 dx_2 dx_3 \cdots dx_N \leq \varepsilon \\
&\qquad\qquad \forall \lambda \geq \lambda_0 \quad \forall t \in (0, t_0).
\end{aligned}
$$

Therefore, in order to prove (3.9), we need only the fact that under the same conditions,

$$
\int_{|x_1|\geq k_0}\int_{\mathbf{R}^{N-1}} u_\lambda(t, x_1, \dots, x_N)dx_1 \cdots dx_N \leq \varepsilon \quad \forall \lambda \geq \lambda_0 \quad \forall t \in (0, t_0), \tag{3.10}
$$

$$
\int_{|x_2|\geq k_0}\int_{\mathbf{R}^{N-1}} u_\lambda(t, x_1, \cdots, x_N)dx_1 \cdots dx_N \leq \varepsilon \quad \forall \lambda \geq \lambda_0 \quad \forall t \in (0, t_0). \tag{3.11}
$$

Let us estimate (3.10); the bound (3.11) can be obtained in the same way. We define

$$v_\lambda(t,x_1,\overline{x}) = \int_{-\infty}^{x_1} u_\lambda(t,z,\overline{x})dz \quad \forall (x_1,\overline{x}) \in \mathbf{R}^N.$$

Then v_λ satisfies

$$\frac{\partial v_\lambda}{\partial t} - \sum_{j=2}^N \frac{\partial^2 v_\lambda}{\partial x_j^2} - \lambda^{1-2\beta}\frac{\partial^2 v_\lambda}{\partial x_1^2} - \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \frac{\partial}{\partial x_2}\int_{-\infty}^{x_1} |u_\lambda|^{p-1}u_\lambda dz = -|u_\lambda|^{q-1}u_\lambda \le 0.$$

Take $w_\lambda(t,x_1) = \int_{\mathbf{R}^{N-1}} v_\lambda(t,x_1,\overline{x})d\overline{x}$. The function w_λ then satisfies

$$\frac{\partial w_\lambda}{\partial t} - \lambda^{1-2\beta}\frac{\partial^2 w_\lambda}{\partial {x_1}^2} \le 0; \qquad w_\lambda(0) = \int_{\mathbf{R}^{N-1}}\int_{-\infty}^{x_1} u_\lambda(t,z,\overline{x})dzd\overline{x},$$

and it is bounded from above by the solution W_λ of the equation

$$\partial_t W_\lambda - \lambda^{1-2\beta}\frac{\partial^2 W_\lambda}{\partial x_1^2} = 0$$

with the same initial data. We have $W_\lambda(t,x_1) = (K_\lambda(t) * w_\lambda(0))(x_1)$ with $K_\lambda(t,x_1) = \lambda^{\beta-1/2}K_1(t,\lambda^{\beta-1/2}x_1)$ and where K_1 is the one-dimensional heat kernel. Let us show that for every $k_0 > 0$ and every $t > 0$,

$$\lim_{\lambda\to\infty} w_\lambda(t,-k_0) = \lim_{\lambda\to\infty}\int_{\mathbf{R}^{N-1}}\int_{-\infty}^{-k_0} u_\lambda(t,z,\overline{x})dx_1 d\overline{x} = 0. \tag{3.12}$$

By definition,

$$\begin{aligned} W_\lambda(t,-k_0) &= \lambda^{\beta-\frac{1}{2}}(4\pi t)^{-\frac{1}{2}}\int_{\mathbf{R}} \exp\left(-\frac{|z|^2}{4t\lambda^{1-2\beta}}\right) w_\lambda(0,-k_0-z)dz \\ &\le M\lambda^{\beta-\frac{1}{2}}(4\pi t)^{-\frac{1}{2}}\int_{\{|z|\ge k/2\}} \exp\left(-\frac{|z|^2}{4t\lambda^{1-2\beta}}\right)dz \\ &\quad + \lambda^{\beta-\frac{1}{2}}(4\pi t)^{-\frac{1}{2}}\int_{\{|z|<k/2\}} \exp\left(-\frac{|z|^2}{4t\lambda^{1-2\beta}}\right) w_\lambda(0,-k_0-z)dz \end{aligned}$$

and

$$\lim_{\lambda\to\infty}\int_{\{|z|\ge k/2\}} \exp\left(-\frac{|z|^2}{4t\lambda^{1-2\beta}}\right)dz = 0$$

since $1-2\beta < 0$. Moreover, by construction, we have uniformly on the set $\{|z| < k/2\}$ that

$$\begin{aligned} \lim_{\lambda\to\infty} w_\lambda(0,-k_0-z) &= \lim_{\lambda\to\infty}\int_{\mathbf{R}^{N-1}}\int_{-\infty}^{-k_0-z} u_\lambda(0,y,\overline{x})dyd\overline{x} \\ &\le \lim_{\lambda\to\infty}\int_{\mathbf{R}^{N-1}}\int_{-\infty}^{-\lambda^\beta k_0/2} u_0(y,\overline{x})dyd\overline{x} = 0, \end{aligned}$$

from which, using that $\int_{\mathbf{R}} K_\lambda(t,y)dy = 1$ for every t, we deduce that

$$\lim_{\lambda\to\infty}\int_{\{|z|<k/2\}} \exp\left(-\frac{|z|^2}{4t\lambda^{1-2\beta}}\right) w_\lambda(0,-k_0-z)dz = 0,$$

and (3.12) follows. It is easy to see that the limits above are uniform on intervals of the form $(0,t_0)$ with $t_0>0$ finite.

On the other hand, now define $w_\lambda(t,x_1)=\int_{\mathbf{R}^{N-1}}\int_{x_1}^{\infty}u_\lambda(t,z,\overline{x})dzd\overline{x}$. Then w_λ satisfies

$$\begin{aligned}\frac{\partial w_\lambda}{\partial t}-\lambda^{1-2\beta}\frac{\partial^2 w_\lambda}{\partial {x_1}^2}&=\int_{\mathbf{R}^{N-1}}|u_\lambda|^{q-1}u_\lambda d\overline{x}\\&\le Ct^{-\frac{N+1}{2q}(q-1)}\int_{\mathbf{R}^{N-1}}u_\lambda(x_1,\overline{x})d\overline{x}\equiv -Ct^{-\frac{N+1}{2q}(q-1)}\frac{\partial w_\lambda}{\partial x_1}.\end{aligned}$$

By the classical parabolic comparison principle, for every $\lambda>0$, the function w_λ is bounded from above by W_λ, the solution of

$$\begin{aligned}\frac{\partial W_\lambda}{\partial t}-\lambda^{1-2\beta}\frac{\partial^2 W_\lambda}{\partial {x_1}^2}+Ct^{-\frac{N+1}{2q}(q-1)}\frac{\partial W_\lambda}{\partial x_1}&=0,\\W_\lambda(0,x_1)=\int_{\mathbf{R}^{N-1}}\int_{x_1}^{\infty}u_\lambda(0,z,\overline{x})dzd\overline{x}&.\end{aligned}$$

Therefore, the function

$$\omega_\lambda(t,x_1)\equiv W_\lambda\left(t,x_1+\frac{C}{\beta}t^\beta\right)$$

(where, remember, β is defined in (2.3)) satisfies

$$\frac{\partial \omega_\lambda}{\partial t}-\lambda^{1-2\beta}\frac{\partial^2\omega_\lambda}{\partial {x_1}^2}=0;\qquad \omega_\lambda(0,x_1)=\int_{\mathbf{R}^{N-1}}\int_{x_1}^{\infty}u_\lambda(0,z,\overline{x})dzd\overline{x}.$$

By the same argument as before, we deduce that for every $k>0$ and for all $t_0>0$,

$$\lim_{\lambda\to\infty}\int_{\mathbf{R}^{N-1}}\int_{k}^{\infty}u_\lambda(t,z,\overline{x})dx_1d\overline{x}=0\quad\forall t\in(0,t_0).\tag{3.13}$$

From (3.12) and (3.13), we deduce (3.10). The same argument gives the estimate for (3.11), and this completes the proof of Lemma 3.4 for the case where $u_0\ge 0$. Of course the same proof holds for nonpositive initial data $u_0\le 0$.

In the general case, let

$$u_0^+=\max(0,u_0),\qquad u_0^-=\min(0,u_0)$$

so that $u_0\equiv u_0^++u_0^-$. Let v and w be the solutions of (0.10) and (0.2) with, respectively, u_0^+ and u_0^- as initial data. By the maximum principle, $w\le u\le v$ on $(0,\infty)\times\mathbf{R}^N$. Thus for every $\lambda>0$, $w_\lambda\le u_\lambda\le v_\lambda$ on $(0,\infty)\times\mathbf{R}^N$. Therefore, using (3.9) for nonpositive and nonnegative initial data, we deduce the lemma in the general case. □

We can show now that the function U takes $M\delta$ as initial data in the sense of (1.7).

Lemma 3.5. *For every function $\varphi\in BC(\mathbf{R}^N)$,*

$$\lim_{t\to 0}\int_{\mathbf{R}^N}U(t,x)\varphi(x)dx=M\varphi(0).\tag{3.14}$$

Proof. Let us prove (3.14) for $\varphi \in \mathcal{D}(\mathbf{R}^N)$ first. To this end, we multiply equation (2.4) by φ and integrate by parts on $(0,t) \times \mathbf{R}^N$. This gives

$$
\begin{aligned}
&\left| \int_{\mathbf{R}^N} u_\lambda(t,x)\varphi(x)dx - \int_{\mathbf{R}^N} u_\lambda(0,x)\varphi(x)dx \right| \\
&\quad \leq \left| \int_0^t \int_{\mathbf{R}^N} u_\lambda(s,x)\widetilde{\triangle}\varphi(x)dxds \right| + \lambda^{1-2\beta}| \int_0^t \int_{\mathbf{R}^N} u_\lambda(s,x)\frac{\partial^2\varphi}{\partial x_1^2}(x)dxds| \\
&\qquad\qquad + \left| \int_0^t \int_{\mathbf{R}^N} |u_\lambda(s,x)|^{q-1} u_\lambda(s,x)\frac{\partial\varphi}{\partial x_1}dxds \right| \\
&\qquad\qquad + \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \left| \int_0^t \int_{\mathbf{R}^N} |u_\lambda(s,x)|^{p-1} u_\lambda(s,x)\frac{\partial\varphi}{\partial x_2}dxds \right| \\
&\quad \leq Mt||\widetilde{\triangle}\varphi||_{L^\infty(\mathbf{R}^N)} + \lambda^{1-2\beta}t\left\| \frac{\partial^2\varphi}{\partial x_1^2} \right\|_{L^\infty(\mathbf{R}^N)} + 2\sqrt{t}\left\| \frac{\partial\varphi}{\partial x_1} \right\|_{L^\infty(\mathbf{R}^N)} \\
&\qquad + \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \frac{t^{1-\frac{N}{2}(p-1)}}{1-\frac{N}{2}(p-1)} \left\| \frac{\partial\varphi}{\partial x_2} \right\|_{L^\infty(\mathbf{R}^N)}.
\end{aligned}
$$

Since $\{u_\lambda(0)\}$ is an approximation of the identity, for every $\varepsilon > 0$, there is a $\tau > 0$ and $\lambda_0 > 0$ such that if $t \in (0,\tau)$ and $\lambda > \lambda_0$,

$$\left| \int_{\mathbf{R}^N} u_\lambda(t,x)\varphi(x)dx - M\varphi(0) \right| \leq \varepsilon.$$

Passing to the limit as λ_k goes to ∞, we obtain

$$\left| \int_{\mathbf{R}^N} U(t,x)\varphi(x)dx - M\varphi(0) \right| \leq \varepsilon \quad \forall t \in (0,\tau).$$

Using Lemma 3.4 and (3.14), we easily conclude the proof of Lemma 3.5. □

From Lemma 3.5, we easily deduce the following for U.

COROLLARY 3.6. *The function U satisfies $U \in \mathbf{C}((0,\infty); L^p(\mathbf{R}^N))$ for every $p \in [1,\infty)$ and $u_{\lambda_k}(t,.) \to U(t,.)$ in $L^p(\mathbf{R}^N)$ uniformly in $t \in [t_1,t_2]$ with $0 < t_1 < t_2 < \infty$. Moreover, it is an entropy solution of* (1.4)–(1.7).

We then deduce by the uniqueness of the function v_M of Proposition 1.2 that $U \equiv v_M$ and that it is the entire family u_λ which converges to v_M as $\lambda \to \infty$ uniformly in $C([t_1,t_2], L^r(\mathbf{R}^N))$ for every $r \in [1,\infty)$ and $0 < t_1 < t_2 < \infty$. As in section 2, this implies that for all such r's,

$$\lim_{\lambda\to\infty} ||u_\lambda(1,\cdot) - v_M(1,\cdot)||_{L^r(\mathbf{R}^N)} = 0. \tag{3.15}$$

By a simple change of variables, (3.15) proves the main result if the initial data $u_0 \geq 0$ satisfy (2.8).

Suppose now that we only have $u_0 \geq 0$ with $u_0 \in L^1(\mathbf{R}^N)$. By (1.1), we have that for any fixed $\tau > 0$, $u(\tau)$ satisfies (2.8). Applying the previous arguments to the function $\widetilde{u}(t) = u(t+\tau)$, we deduce the main result for nonnegative initial data $u_0 \in L^1(\mathbf{R}^N)$.

4. Proof of the main result for general initial data. In this section, we consider the solutions of (0.10) and (0.2) where the initial data u_0 may be any function of $L^1(\mathbf{R}^N)$, not necessarily positive or negative but which may change sign. Then we no longer have the entropy inequality (2.9) anymore. Regardless, since

$$\forall t > 0, \quad \forall x \in \mathbf{R}^N, \quad |u(t,x)| \leq \overline{u}(t,x),$$

where $\overline{u}$ is the solution of (0.10) with initial data $\overline{u}(0) = |u_0|$, the decay estimate (2.10) is still true. The proof follows the same lines as in the case of nonnegative initial data. The only difference is how we obtain the compactness of the family $\{u_\lambda\}$. Since we cannot use the entropy inequality, we use a general compactness result of [LPT] based on the kinetic approach and slightly modified in [C1] and [C2] to handle nonlinearities which are not of class C^2.

Proposition 4.1. *Let $\rho_\lambda \in C(0,\infty; L^1(\mathbf{R}^N)) \cap L^\infty(\mathbf{R}^N \times (0,\infty))$ be a family of solutions of*

$$\frac{\partial \rho_\lambda}{\partial t} - \lambda^{1-2\beta}\frac{\partial^2 \rho_\lambda}{\partial x_1^2} - \sum_{j=2}^{N} \frac{\partial^2 \rho_\lambda}{\partial x_j^2} + \frac{\partial |\rho_\lambda|^{q-1}\rho_\lambda}{\partial x_1} + \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \frac{\partial |\rho_\lambda|^{p-1}\rho_\lambda}{\partial x_2} = 0 \tag{4.1}$$

such that the following hold:

1. *For all convex functions S,*

(4.2)
$$\frac{\partial S(\rho_\lambda)}{\partial t} - \lambda^{1-2\beta}\frac{\partial^2 \eta_{11}(\rho_\lambda)}{\partial x_1^2} - \sum_{j=2}^{N} \frac{\partial^2 \eta_{jj}(\rho_\lambda)}{\partial x_j^2} + \frac{\partial \eta_1(\rho_\lambda)}{\partial x_1} + \lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \frac{\partial \eta_2(\rho_\lambda)}{\partial x_2} \leq 0,$$

where

$$\eta_1(t) = q\int_0^t S'(s)|s|^{q-1}ds, \qquad \eta_2(t) = p\lambda^{\frac{1}{2}-\frac{(N+1)(p-1)}{2q}} \int_0^t S'(s)|s|^{p-1}ds,$$
$$\eta_{11} = \lambda^{1-2\beta}(S(t) - S(0)), \qquad \eta_{jj} = (S(t) - S(0)) \quad \forall j \in \{2, \ldots, N\}.$$

2. *ρ_λ is bounded in $L^\infty((0,\infty) \times \mathbf{R}^N) \cap L^\infty((0,\infty); L^1(\mathbf{R}^N))$ uniformly in λ.*

Then ρ_λ is relatively compact in $L^1_{\text{loc}}((0,\infty) \times \mathbf{R}^N)$.

By (2.10), the family u_λ satisfies condition 2 above. Since for every $\lambda > 0$, the function u_λ is a classical solution of the parabolic equation (4.1), it satisfies (4.2). We then deduce that u_λ is relatively compact in $L^1_{\text{loc}}((0,\infty) \times \mathbf{R}^N)$. The proof then follows the same lines as in section 3.

5. Further comments.

5.1. More general nonlinearities. With some trivial modifications in the proof, one can show that the main result is still true for the solutions of the Cauchy problem with initial data $u_0 \in L^1(\mathbf{R}^N)$ that satisfies

$$u_t - \triangle u + \sum_{i=1}^{N} \frac{\partial |u|^{r_i - 1}u}{\partial x_i} = 0 \quad \text{in } Q = \mathbf{R}^N \times (0,\infty) \tag{5.1}$$

with $\{r_i\}_{i=1}^N$ such that for some $q \in (1, 1 + 1/N)$, either $r_i = q$ or $r_i > 1 + q/(N+1)$.

Under slight modifications in the proof, the same result can be extended to small perturbations of the pure power-like case. More precisely, consider the Cauchy problem with initial data in $L^1(\mathbf{R}^N)$ associated with the equation

$$u_t - \triangle u + \sum_{i=1}^{N} \frac{\partial f_i(u)}{\partial x_i} = 0 \quad \text{in } (0,\infty) \times \mathbf{R}^N \tag{5.2}$$

with $f_i \in C^1([0,\infty)) \cap C^2((0,\infty))$, $f_i(0) = f_i'(0) = 0$. Moreover, suppose that there exist $q \in (1, 1+1/N)$, $j \in \{1, \ldots, N\}$, and $r_i > 1 + q/(N+1)$ for $i \neq j$ such that the limits

$$\lim_{s \to 0} \frac{f_j''(s)}{|s|^{q-3}s} = q(q-1)C \neq 0 \quad \text{and} \quad \lim_{s \to 0} \frac{f_i''(s)}{|s|^{r_i - 3}s}, \quad i \neq j,$$

exist. Then the main result stated in the introduction remains true for the solutions of (5.2) and (0.2), where the limiting equation is

$$u_t - \tilde{\triangle} u + C\frac{\partial |u|^{q-1}u}{\partial x_j} = 0.$$

5.2. Open problems. We cannot consider the case where one of the powers r_i is such that $q < r_i < 1 + q/(N+1)$. Even the simplest equation

$$u_t - \triangle u + \frac{\partial |u|^{q-1}u}{\partial x_1} + \frac{\partial |u|^{p-1}u}{\partial x_2} = 0$$

with $1 < q < p \leq 1 + q/(N+1)$ is out of the scope of our results.

However, looking at the scaling transformation (2.4), we see that when $p = 1 + q/(N+1)$, the nonlinearities remain invariant. Therefore, in this particular case, one expects the large-time behavior to be given by self-similar entropy solutions of the reduced equation

$$u_t - \sum_{j=2}^{N} \frac{\partial^2 u}{\partial {x_j}^2} + \frac{\partial |u|^{q-1}u}{\partial x_1} + \frac{\partial |u|^{p-1}u}{\partial x_2} = 0.$$

Note that the two nonlinearities are present. The large-time behavior is probably of a completely different nature in the range $1 < q < p < 1 + q/(N+1)$.

REFERENCES

[AEZ] J. Aguirre, M. Escobedo, and E. Zuazua, *Self-similar solutions of a convection diffusion equation and related semilinear problems,* Comm. Partial Differential Equations, 15 (1990), pp. 139–157.

[B] H. Brezis, *Analyse Fonctionnelle: Théorie et Applications,* Masson, Paris, 1983.

[C1] A. Carpio, *Unicité et comportement asymptotique pour des équations de convection-diffusion scalaires,* C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 51–56.

[C2] A. Carpio, *Large time behavior in convection-diffusion equations,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), to appear.

[EVZ1] M. Escobedo, J. L. Vázquez, and E. Zuazua, *Source-type solutions and asymptotic behaviour for a diffusion-convection equation,* Arch. Rational Mech. Anal., 124 (1993), pp. 43–66.

[EVZ2] M. Escobedo, J. L. Vázquez, and E. Zuazua, *A diffusion-convection equation in several space dimensions*, Indiana Univ. Math. J., 42 (1993), pp. 1413–1440.

[EVZ3] M. Escobedo, J. L. Vázquez, and E. Zuazua, *Entropy solutions for diffusion-convection equations with partial diffusivity,* Trans. Amer. Math. Soc., 343 (1994), pp. 829–842.

[EZ] M. Escobedo and E. Zuazua, *Large-time behaviour for solutions of a convection diffusion equation in* $\mathbf{R}^N$*,* J. Funct. Anal., 100 (1991), pp. 119–161.

[K] S. Kawashima, *Self-similar solutions of a convection-diffusion equation,* in Nonlinear PDE: Japan Symposium 2, Lecture Notes in Numer. Appl. Anal. 12, K. Masuda, M. Mimura, and T. Nishida, eds., Springer-Verlag, Berlin, 1993, pp. 123–136.

[Kr] S. Kruzhkov, *First-order quasilinear equations in several independent variables,* Math. USSR-Sb., 10 (1970), pp. 217–243.

[LPT] P.L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations,* J. Amer. Math. Soc., 7 (1994), pp. 169–189.

[S] J. Simon, *Compact sets in the space* $L^p(0,T;B)$*,* Ann. Mat. Pura Appl., CXLVI (1987), pp. 65–96.

[Z1] E. Zuazua, *A dynamical system approach to the self similar large time behavior in scalar convection-diffusion equations,* J. Differential Equations, 108 (1994), pp. 1–35.

[Z2] E. Zuazua, *Weakly nonlinear large time behavior for scalar convection-diffusion equations,* Differential Integral Equations, 6 (1993), pp. 1481–1492.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 595–610, May 1997

006

STABILITY AND LYAPUNOV FUNCTIONS FOR REACTION-DIFFUSION SYSTEMS[*]

W. B. FITZGIBBON[†], S. L. HOLLIS[‡], AND J. J. MORGAN[§]

Abstract. It is shown for a large class of reaction-diffusion systems with Neumann boundary conditions that in the presence of a separable Lyapunov structure, the existence of an a priori L^r estimate, uniform in time, for some $r > 0$, implies the L^∞-uniform stability of steady states. The results are applied to a general class of Lotka–Volterra systems and are seen to provide a partial answer to the global existence question for a large class of balanced systems with nonlinearities that are not bounded by any polynomial.

Key words. reaction-diffusion systems, stability, Lyapunov functions

AMS subject classifications. 35B35, 35K45, 35K57

PII. S0036141094272241

1. Introduction.

One of the persistent problems in the theory of systems of reaction-diffusion equations concerns the description of the qualitative effects of adding diffusion to systems of ordinary differential equations. To be more precise, if $f = (f_i)_{i=1}^m\colon \mathbb{R}^m \to \mathbb{R}^m$, then solutions to the system of ordinary differential equations

$$\begin{aligned} &\dot{u}(t) = f(u(t)), \quad t > 0,\\ &u(0) = u_0 \end{aligned} \tag{1.1}$$

determine constant solutions to the reaction-diffusion system

$$\begin{aligned} &\partial u/\partial t = D\Delta u + f(u) && \text{on } \Omega \times (0,\infty),\\ &\partial u/\partial \mathbf{n} = 0 && \text{on } \partial\Omega \times (0,\infty),\\ &u(\,\cdot\,,0) = u_0(\,\cdot\,) && \text{on } \Omega, \end{aligned} \tag{1.2}$$

where $u = (u_1, \ldots, u_m)^T$, D is a diagonal matrix with distinct entries $d_i > 0$ along the diagonal and Δ denotes the vector Laplacian. One principal question associated with these systems is whether or not global existence of solutions to (1.1) for all choices of initial data guarantees global existence of solutions to (1.2) for all choices of sufficiently smooth initial data. This question remained unresolved until the recent work of Pierre and Schmidt [19]. In that work, the authors give an example of a two-component system for which solutions to (1.1) exist globally while those to the partial differential equation blow up in finite time. Their work is related to a long-standing question pointed out by Martin in the early 1980s regarding the global existence of nonnegative solutions to two component systems of the form

$$\begin{aligned} &\partial u/\partial t = d_1\Delta u + f_1(u,v),\\ &\partial v/\partial t = d_2\Delta v + f_2(u,v) \end{aligned} \quad \text{on } \Omega \times (0,\infty), \tag{1.3}$$

[*] Received by the editors August 1 1994; accepted for publication (in revised form) January 16, 1996.
http://www.siam.org/journals/sima/28-3/27224.html

[†] Department of Mathematics, University of Houston, Houston, TX 77204 (fitz@math.uh.edu). The research of this author was supported in part by NSF grant DMS-9207064.

[‡] Department of Mathematics and Computer Science, Armstrong Atlantic State University, Savannah, GA 31419 (selwyn_hollis@mailgate.armstrong.edu).

[§] Department of Mathematics, Texas A&M University, College Station, TX 77843 (jmorgan@math.tamu.edu). The research of this author was supported in part by NSF grant DMS-9208046.

where $f_1(0,v)$, $f_2(u,0) \geq 0$ for all $u, v \geq 0$ and $f_1(u,v) + f_2(u,v) \leq 0$ for all $u, v \geq 0$. These two conditions on the vector field (f_1, f_2) are referred to, respectively, as quasi positivity and balancing.

This has given rise to similar questions about more general, balanced, quasi-positive reaction-diffusion systems of the form

$$(1.4) \qquad \partial u_i/\partial t = d_i \Delta u_i + f_i(u) \quad \text{on } \Omega \times (0,\infty), \quad i = 1, \dots, m,$$

where $u = (u_1, \dots, u_m)^T$, $f_i(u) \geq 0$ whenever $u \in \mathbb{R}^m_+$ with $u_i = 0$, and $\sum_{i=1}^m f_i(u) \leq 0$ for all $u \in \mathbb{R}^m_+$. It should be noted that these balancing and quasi positivity assumptions easily imply that $f(0) = 0$ and that all solutions of the ordinary differential equation in (1.1) having nonnegative initial data exist and are bounded for all $t \geq 0$. In particular, for any $M > 0$, the region

$$\left\{ u \;\Big|\; \sum_{i=1}^n u_i \leq M, \quad u_i \geq 0 \right\}$$

is invariant for these systems; therefore, the zero solution is stable with respect to $\mathbb{R}^m_+$.

This structure is merely a simple case of a more general, separable Lyapunov structure for (1.1). Such a structure has the form $H(u) = \sum_{i=1}^m h_i(u_i)$, where H: $\mathbb{R}^m_+ \to [0,\infty)$ is a convex function that has a unique zero in $\mathbb{R}^m_+$ and whose level hypersurfaces bound invariant regions for solutions of (1.1). The existence of such an H easily guarantees the stability of the steady state z. Recent work in this vein includes [1], [5], [6], [10], [14], and [16].

The work at hand concerns the persistence of stability of steady-state solutions to (1.1) in the presence of a separable Lyapunov structure when diffusion is added to the system, that is, in the setting of (1.2) with nonnegative, continuous initial data. Questions of stability for nonlinear systems are frequently resolved via linearized stability or Lyapunov-type methods. Typically, when one attempts to lift Lyapunov functions from the setting of (1.1) to (1.2), one obtains estimates in $L_1(\Omega)$ or $L_p(\Omega)$ and not the optimal uniform $L_\infty(\Omega)$ estimates needed to obtain stability. Therefore, the central theme of our work will be the introduction of an intermediate notion of stability from $C(\overline{\Omega})$ to $L_p(\Omega)$, and the bootstrapping of $L_p(\Omega)$ estimates to $L_\infty(\Omega)$.

We should point out the phenomena of diffusion driven instabilities. It is well known that the addition of diffusion can destabilize constant steady states; see, e.g., [3] and [18]. Therefore, we shall be lead to the conclusion that the systems of ordinary differential equations which admit diffusion-driven instability do not have a Lyapunov structure of the type to be described.

Our subsequent development consists of five sections. In addition to detailing our hypotheses and outlining relevant theory, the second section introduces the central notion of stability from $C(\overline{\Omega})$ to $L_p(\Omega)$ and bootstraps this stability from $L_p(\Omega)$ to $L_\infty(\Omega)$. As such, the second section forms the theoretical basis of the paper. The third section introduces the notion of D-diffusively convex Lyapunov functionals and demonstrates the connection to the work in section 2. The fourth section is concerned with application of the theory. It begins by considering balanced two-component systems and then applies the theory to dissipative chemical systems and Lotka–Volterra systems. We conclude with some general comments and remarks.

We conclude this section with two remarks. First, our result gives a partial answer to Martin's original question. We determine that balanced, quasi-positive reaction-diffusion systems subject to homogeneous Neumann boundary conditions have global solutions for all choices of continuous, sufficiently small, nonnegative initial data. Second, we have limited our discussion to the case of homogeneous Neumann boundary conditions for the following reason. If all components of our system satisfy strictly dissipative boundary conditions, such as homogeneous Dirichlet or homogeneous Robin, then the presence of a separable Lyapunov structure along with these boundary conditions allows one to employ linearized stability arguments to obtain asymptotic stability. In the case of a mixture of homogeneous Neumann boundary conditions and dissipative conditions (as mentioned above), the arguments follow our development.

2. Preliminaries and ∞–r stability. In what follows, Ω shall be a bounded domain in $\mathbb{R}^n$ that lies locally on one side of its $C^{2+\alpha}$ boundary $\partial\Omega$. We shall always assume that the initial data $u_0 = (u_{0_1}, \ldots, u_{0_m})^T \in C(\Omega)^m$ and that the vector field $f = (f_i)_{i=1}^m$ has the property that

$$f \in C^1(\mathbb{R}^m; \mathbb{R}^m). \tag{2.1}$$

However, we make no assumptions concerning the growth rates of the individual components f_i of f. The symbol D will denote an $m \times m$ diagonal matrix with distinct entries $d_i > 0$, $i = 1$ to m, along the diagonal. We point out that all results contained herein would trivialize were we to assume that the d_i's are identical. We hope that we shall not introduce undue confusion by using the symbol "Δ" to denote both the vector and the scalar Laplacian. Equations without subscripts will typically denote vector equations and nonsubscripted scalar equations shall be specifically referred to as such.

In our general discussion, we use the notation $z_0 = (z_{0_1}, \ldots, z_{0_m})^T \in \mathbb{R}^m$ to denote an equilibrium point (or steady state) of (1.1). Namely, we have

$$f(z_0) = 0. \tag{2.2}$$

A closed subset of $M \subseteq \mathbb{R}^m$ will be called a forward invariant set for (1.2) if $u_0(x) = (u_{0_1}(x), \ldots, u_{0_m}(x))^T \in M$ for all $x \in \Omega$ implies that

$$u(x,t) = (u_1(x,t), \ldots, u_m(x,t)) \in M \tag{2.3}$$

for all $(x,t) \in \Omega \times [0, T_{\max})$. Here $[0, T_{\max})$ denotes the maximal interval of existence for solutions to the initial boundary value problem (1.2). We shall require that there exists a forward invariant set M (not necessarily bounded) for solutions to (1.2). Hence because the d_i's are assumed to be distinct, we assume that there exists a forward invariant m-cube

$$M = M_1 \times \cdots \times M_m \tag{2.4}$$

for (1.2), where each M_i, $i = 1$ to m, is a closed interval. *We point out that we have said nothing concerning the boundedness of M, and consequently we make no presuppositions concerning the global existence of solutions to* (1.2). For example, M may well be $\mathbb{R}^m_+$ (the positive orthant) or all of $\mathbb{R}^m$.

In what follows, the mild abuse of notation $v \in M$ will be used frequently to indicate that a function $v : \Omega \to \mathbb{R}^m$ has the property that $v(x) \in M$ for all $x \in \Omega$.

Our analysis will involve the standard Lebesgue spaces $L_p(\Omega)$, $p \geq 1$:

$$L_p(\Omega) = \left\{ u \;\middle|\; \int_\Omega |u|^p dx < \infty \right\}; \tag{2.5}$$

$$\|u\|_{p,\Omega} = \left(\int_\Omega |u|^p dx \right)^{1/p}. \tag{2.6}$$

We shall also want to consider the analogous spaces obtained with $0 < p < 1$. Although (2.6) does not define a norm on $L_p(\Omega)$ if $0 < p < 1$, we will use the same notation for the functional defined on $L_p(\Omega)$ by the right side of (2.6). If $p \geq 1$ and $k > 0$, then $W_p^{(k)}(\Omega)$ denotes the usual kth-order Sobolev space in $L_p(\Omega)$ and $W_p^{(2k,k)}(\Omega \times (\tau, T))$ denotes its analogue in $L_p(\Omega \times (\tau, T))$. For definitions of these spaces for both integral and nonintegral k, we refer the reader to [12].

We will need the following fractional-Sobolev-space embedding theorem of Amann [2].

Theorem 2.1. *Let $k \in \mathbb{N}$ and suppose that $\partial\Omega$ is uniformly regular of class C^k. If $0 \leq s' \leq s \leq k$ and $1 < p, q < \infty$, then $W_p^s(\Omega)$ embeds continuously in $W_q^{s'}(\Omega)$ whenever $1/p \geq 1/q$ and $s - (n/p) \geq s' - (n/q)$.*

We now introduce the notion of ∞–r stability. It will be a notion of stability with respect to M, which will allow us to consider steady states belonging to ∂M.

Definition 2.2. *Let $z_0 \in M$ be an equilibrium point of the vector field $f = (f_i)_{i=1}^m$ and let $0 < r \leq \infty$. Then z_0 is said to be uniformly ∞–r stable with respect to M if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $u_0 \in M$ and $\|u_{0_i} - z_{0_i}\|_{\infty,\Omega} < \delta$ for $i = 1$ to m imply*

(i) *a classical solution to (1.2) exists on $\Omega \times [0, \infty)$;*

(ii) *$\|u_{0_i}(\cdot\,, t) - z_{0_i}\|_{r,\Omega} < \varepsilon$ for $i = 1$ to m and $t > 0$.*

An ∞–r stable equilibrium point $z_0 \in M$ is said to be uniformly ∞–r asymptotically stable if there exists a $\delta > 0$ such that $u_0 \in M$ and $\|u_i - z_{0_i}\|_{\infty,\Omega} < \delta$ for $i = 1$ to m imply

(iii) *$\lim_{t\to\infty} \|u_i(\cdot\,, t) - z_{0_i}\|_{r,\Omega} = 0$ for $i = 1$ to m.*

The usual notions of stability with respect to M now correspond to ∞–∞ stability with respect to M as stated formally in the following definition.

Definition 2.3. *An equilibrium point $z_0 \in M$ is said to be uniformly stable with respect to M if it is uniformly ∞–∞ stable with respect to M. A stable equilibrium point $z_0 \in M$ is said to be uniformly asymptotically stable with respect to M if it is uniformly ∞–∞ asymptotically stable with respect to M.*

We shall see in what follows that the notion of ∞–r stability is intermediate and may be subsumed by the notion of stability. For a given point $v_0 \in \mathbb{R}^m$, let the symbol $C_\eta(v_0)$ denote the m-dimensional cube centered at v_0 with diameter $2\sqrt{m}\,\eta$ and $B_\delta(v_0)$ denote the m-dimensional ball of radius δ about v_0. We remark that $C_\varepsilon(v_0) \subseteq B_{\sqrt{m}\,\varepsilon}(v_0)$. The analysis that follows will require "cutoff" functions $\varphi_{\eta,v_0} \in C^\infty(\mathbb{R}^m; [0, 1])$, defined for $\eta > 0$ by

$$\begin{aligned} &\varphi_{\eta,v_0}(u) = 1 \quad \text{for } u \in C_\eta(v_0), \\ &\varphi_{\eta,v_0}(u) = 0 \quad \text{for } u \in \mathbb{R}^m \setminus C_{2\eta}(v_0). \end{aligned} \tag{2.7}$$

If $z_0 \in M$ is an equilibrium point, we truncate the vector field f by componentwise multiplication by φ_{η,z_0} for $\eta > 0$, i.e., we define $f[\eta, z_0] = (f_i[\eta, z_0])_{i=1}^m$ by

$$f_i[\eta, z_0](u) = \varphi_{\eta,z_0}(u) f_i(u). \tag{2.8}$$

Then solutions to the truncated system

$$\partial v/\partial t = D\Delta v + f[\eta, z_0](v) \quad \text{on } \Omega \times (0, \infty), \tag{2.9a}$$

$$\partial v/\partial \mathbf{n} = 0 \quad \text{on } \partial\Omega \times (0, \infty), \tag{2.9b}$$

$$v(\cdot\,, 0) = v_0 \quad \text{on } \Omega, \tag{2.9c}$$

where $v_{0_i} = \varphi_{\eta,z_0}(u_0)u_{0_i}$, exist on $\Omega \times [0, \infty)$ and are globally bounded. Moreover, it is trivial to observe that if $z_0 \in M$ is an equilibrium point of f, then z_0 is also an equilibrium point of $f[\eta, z_0]$.

We now formally state a few simple observations concerning solutions to (2.9).

LEMMA 2.4. *If $\eta > 0$, $v_0 \in C(\overline{\Omega}; C_{2\eta}(z_0))$, and $f[\eta, z_0]$ is the vector field defined via* (2.8)*, then* (2.9) *has a unique classical solution on $\Omega \times [0, \infty)$. Moreover,*

(i) *$v(\cdot, t) \in C_{2\eta}(z_0) \cap M$ for $t \geq 0$;*

(ii) *if $v(\cdot, t) \in C_\eta(z_0) \cap M$ for $0 \leq t < T$, then $v(x, t) = u(x, t)$ for $(x, t) \in \Omega \times [0, T)$, where u is the solution to* (1.2).

Proof. We observe that $M \cap C_{2\eta}(z_0)$ is a bounded invariant region for (2.9) because the vector field $f[\eta, z_0]$ is identically zero exterior to $C_{2\eta}(z_0)$ and does not point out of M. Therefore, solutions to (2.9) exist globally and remain confined to $M \cap C_{2\eta}(z_0)$ for all time [21]. Classical uniqueness theory for parabolic equations together with the observation that $f[\eta, z_0]|_{C_\eta(z_0)} = f|_{C_\eta(z_0)}$ immediately confirms the second assertion. □

LEMMA 2.5. *If z_0 is uniformly ∞–r stable for* (2.9) *with respect to M for some $r \in (0, \infty)$, then z_0 is uniformly ∞–p stable for* (2.9) *with respect to M for all $p \in (0, \infty)$. Analogous results hold for uniform ∞–r asymptotic stability.*

Proof. If $0 < p < r$, the results follow easily by the Jensen inequality and the convexity of $g(z) = |z|^{r/p}$. If $r < p$, then because $g(z) = |z - z_{0_i}|^{p-r}$ is bounded above by $(2\sqrt{m}\,\eta)^{p-r}$ on $C_{2\eta}(z_0)$ and since $v(x, t) \in C_{2\eta}(z_0)$ for all $(x, t) \in \Omega \times (0, \infty)$, we have

$$(2.10) \qquad |v_i - z_{0_i}|^p \leq (2\sqrt{m}\,\eta)^{p-r} |v_i - z_{0_i}|^r \quad \text{on } \Omega \times (0, \infty),$$

from which the desired results follow. □

The next theorem provides the foundation of our development. It states that ∞–r stability of the truncated system (2.9) guarantees ∞–∞ stability of the original system (1.2).

THEOREM 2.6. *Let $z_0 \in M$ be an equilibrium point of the vector field f. If $r > 0$ and z_0 is a uniformly ∞–r stable equilibrium point for* (2.9)*, then z_0 is a uniformly stable solution for* (1.2)*. Analogous results hold for uniformly ∞–r asymptotic stability.*

Proof. We begin by fixing $\eta > 0$. If we are able to choose $\delta > 0$ so that solutions to (2.9) have the property that $v_0 \in C_\delta(z_0)$ implies that $v(x, t) \in C_\varepsilon(z_0)$, where $\varepsilon < \eta$, then solutions to (2.9) and (1.2) coincide. Therefore, it will suffice to demonstrate that uniformly ∞–r stable solutions of (2.9) are uniformly stable solutions of (2.9).

By virtue of Lemma 2.5 with $p = 2$, we know that there exists a continuous function $\tilde{\rho}_1$ with $\tilde{\rho}_1(0) = 0$ and $\tilde{\rho}_1(s) > 0$ for $s > 0$ such that for $i = 1$ to m and $t \in [0, \infty)$,

$$(2.11) \qquad \|v_i(\cdot, t) - z_{0_i}\|_{2,\Omega} \leq \tilde{\rho}_1(\|v_0 - z_0\|_{\infty,\Omega}).$$

We shall demonstrate via an iteration scheme that there exists a continuous function ρ with $\rho(0) = 0$ and $\rho(s) > 0$ such that for $i = 1$ to m and $t \in [0, \infty)$, we have

$$(2.12) \qquad \|v_i(\cdot, t) - z_{0_i}\|_{\infty,\Omega} \leq \rho(\|v_0 - z_0\|_{\infty,\Omega}),$$

and we shall thereby obtain our desired conclusion. Toward this end, we set

$$(2.13) \qquad w(x, t) = v(x, t) - z_0$$

and multiply the ith component of (2.9a) by w_i to obtain

$$w_i \partial w_i/\partial t - w_i d_i \Delta w_i = w_i f_i[\eta, z_0](v). \tag{2.14}$$

Because $f[\eta, z_0]$ is Lipschitz, there exists an N such that integration of (2.14) on the space–time cylinder $\Omega \times (\tau, T)$ yields

$$\begin{aligned} &\frac{1}{2}\|w_i(\cdot\,, T)\|_{2,\Omega}^2 + d_i \int_\tau^T \int_\Omega |\nabla w_i|^2 dx dt \\ &\quad\le \frac{1}{2}\|w_i(\cdot\,, \tau)\|_{2,\Omega}^2 + N \sum_{k=1}^m \int_\tau^T \int_\Omega |w_i||w_k| dx dt. \end{aligned}$$

This implies that if $\tau \ge 0$ and $\tau + 1 < T < \tau + 3$, then

$$\begin{aligned} &\frac{1}{2}\|w_i(\cdot\,, T)\|_{2,\Omega}^2 + d_i \int_{\tau+1}^T \int_\Omega |\nabla w_i|^2 dx dt \\ &\quad\le \frac{1}{2}\|w_i(\cdot\,, \tau)\|_{2,\Omega}^2 + N \sum_{k=1}^m \max_{[\tau,\, \tau+3]} \int_\Omega |w_i||w_k| dx. \end{aligned} \tag{2.15}$$

After applying Young's inequality and (2.11) to the right side of (2.15) and the mean-value theorem for integrals to the t-integral on the left side, we construct an increasing sequence $\{T_{1,j}\}_{j=1}^\infty$ with

$$T_{1,1} \le 3 \quad \text{and} \quad 1 < T_{1,j+1} - T_{1,j} < 3 \quad \forall j \in \mathbb{N} \tag{2.16}$$

and a continuous function ρ_1 with $\rho_1(0) = 0$ and $\rho_1(s) > 0$ for $s > 0$ such that

$$\|w_i(\cdot\,, T_{1,j})\|_{2,\Omega}^{(1)} \le \rho_1(\|v_0 - z_0\|_{\infty,\Omega}) \quad \forall j \in \mathbb{N}. \tag{2.17}$$

Now we begin to make use of a well-known classical estimate for parabolic initial boundary value problems from Ladyženskaja, Solonnikov, and Uralćeva [12, p. 341]. More specifically, recall that if $1 < q < \infty$, $0 < \tau < T \le \tau + 3$, $\theta \in L^q(\Omega \times (\tau, T))$, $\phi_0 \in W_q^{2-2/q}(\Omega)$, and ϕ solves

$$\begin{aligned} \partial\phi/\partial t &= d_i \Delta\phi + \theta && \text{on } \Omega \times (\tau, T), \\ \partial\phi/\partial \mathbf{n} &= 0 && \text{on } \partial\Omega \times (\tau, T), \\ \phi(\cdot\,, \tau) &= \phi_0 && \text{on } \Omega, \end{aligned} \tag{2.18}$$

then there exists $c > 0$ such that

$$\|\phi\|_{q,\Omega\times(\tau,T)}^{(2,1)} \le c\Big[\|\theta\|_{q,\Omega\times(\tau,T)} + \|\phi_0\|_{q,\Omega}^{(2-2/q)}\Big], \tag{2.19}$$

where c depends only on d_i and Ω. Applying this parabolic regularity estimate with $q = 2$, we obtain a constant $c_1 > 0$ such that

$$\|w_i\|_{2,\Omega\times(T_{1,j},T_{1,j+1})}^{(2,1)} \le c_1 \Big(\|f_i[\eta, z_0](v)\|_{2,\Omega\times(T_{1,j},T_{1,j+1})} + \|w_i(\cdot\,, T_{1,j})\|_{2,\Omega}^{(1)}\Big). \tag{2.20}$$

We now claim that for every $k \in \mathbb{N}$, there exist

(i) a sequence $\{T_{k,j}\}_{j=1}^\infty$ such that $T_{k,1} \le k+2$ and $1 < T_{k,j+1} - T_{k,j} < 3$ $\forall j \in \mathbb{N}$,

(ii) a constant $c_k > 0$, and

(iii) a function $\rho_k \in C([0,\infty),[0,\infty))$ such that $\rho_k(0)=0$
such that for all $j \in \mathbb{N}$, the estimate

$$\|w_i\|^{(2,1)}_{q_k,\Omega\times(T_{k,j},T_{k,j+1})} \leq c_k\Big(\|f_i[\eta,z_0](v)\|_{q_k,\Omega\times(T_{k,j},T_{k,j+1})} + \rho_k(\|v_0-z_0\|_{\infty,\Omega})\Big) \tag{2.21}$$

is valid with $q_k = 2\big((n+2)/n\big)^{k-1}$.

To establish this claim, we begin by noting that (2.16), (2.17), and (2.20) combine to give the claim for $k=1$. We now proceed by induction on k. Suppose that the claim holds for $k=\ell\geq 1$ and consider the case where $k=\ell+1$. Since $f_i[\eta,z_0]$ is Lipschitz, we can use our hypothesis and Lemma 2.5 with $p=q_\ell$ to conclude from (2.21) that there exists a continuous function $\tilde{\rho}_\ell$ such that $\tilde{\rho}_\ell(0)=0$ and

$$\|w_i\|^{(2,1)}_{q_\ell,\Omega\times(T_{\ell,j},\,T_{\ell,j+4})} \leq \tilde{\rho}_\ell(\|v_0-z_0\|_\infty) \quad \forall j\in\mathbb{N}. \tag{2.22}$$

Note that $T_{\ell,j+4}-T_{\ell,j}>4$. Therefore, (2.22) implies the following inequalities:

$$\int_{T_{\ell,j}}^{T_{\ell,j}+1}\Big(\|w_i\|^{(2)}_{q_\ell,\Omega}\Big)^{q_\ell}dt,\ \int_{T_{\ell,j}+2}^{T_{\ell,j}+3}\Big(\|w_i\|^{(2)}_{q_\ell,\Omega}\Big)^{q_\ell}dt \leq \big[\tilde{\rho}_\ell(\|v_0-z_0\|_{\infty,\Omega})\big]^{q_\ell}. \tag{2.23}$$

Consequently, we can construct a sequence $\{T_{\ell+1,j}\}_{j=1}^\infty$ with

$$T_{\ell,k} < T_{\ell+1,2k-1} < T_{\ell,k}+1 \quad \text{and} \quad T_{\ell,k}+2 < T_{\ell+1,2k} < T_{\ell,k}+3 \tag{2.24}$$

such that

$$\|w_i(\cdot\,,T_{\ell+1,j})\|^{(2)}_{q_\ell,\Omega} \leq \tilde{\rho}_\ell(\|v_0-z_0\|_\infty) \quad \forall j\in\mathbb{N}. \tag{2.25}$$

We now apply Theorem 2.1 to conclude that $W^{(2)}_{q_\ell}(\Omega)$ imbeds continuously into $W^{(2-2/q_{\ell+1})}_{q_{\ell+1}}(\Omega)$. Therefore, there exists $\rho_{\ell+1}\in C([0,\infty),[0,\infty))$ such that $\rho_{\ell+1}(0)=0$ and

$$\|w_i(\cdot\,,T_{\ell+1,j})\|^{(2-2/q_{\ell+1})}_{q_{\ell+1},\Omega} \leq \rho_{\ell+1}(\|v_0-z_0\|_\infty) \quad \forall j\in\mathbb{N}. \tag{2.26}$$

Now by combining (2.26) with the parabolic regularity estimate in (2.19), we see that our claim is true for $k=\ell+1$, thus establishing the claim for all $k\in\mathbb{N}$.

Now with k taken such that $q_k>(n+2)/2$, we have from [12] that there exists $C>0$ such that

$$\|w\|_{\infty,\Omega\times(T_{k,j},T_{k,j+1})} \leq C\|w\|^{(1,2)}_{q_k,\Omega\times(T_{k,j},T_{k,j+1})} \quad \forall j\in\mathbb{N}.$$

Therefore, if we combine this with our claim above, we find that there exists a continuous function $\tilde{\rho}_k$ such that $\tilde{\rho}_k(0)=0$ and

$$\|w\|_{\infty,\Omega\times(T_{k,j},\,T_{k,j+1})} \leq \tilde{\rho}_k(\|v_0-z_0\|_\infty) \quad \forall j\in\mathbb{N}.$$

However, $T_{k,1}\leq k+2$, so

$$\|w\|_{\infty,\Omega\times[k+2,\infty)} \leq \tilde{\rho}_k(\|v_0-z_0\|_\infty). \tag{2.27}$$

We now recall that the operator $-d_i\Delta$ with homogeneous Neumann boundary conditions generates a nonexpansive analytic semigroup $T_i(t)$ on $C(\overline{\Omega})$; see Stewart [22]. Therefore, we have

$$\begin{aligned} w_i(t) &= T_i(t)(v_{0_i} - z_{0_i}) + \int_0^t T_i(t-s) f_i[\eta, z_0](v(\cdot\,, s))ds \\ &= T_i(t)(v_{0_i} - z_{0_i}) + \int_0^t T_i(t-s)\Big(f_i[\eta, z_0](v(\cdot\,, s)) - f_i[\eta, z_0](z_0)\Big)ds, \end{aligned}$$

which implies that

$$\|w_i(t)\|_{\infty,\Omega} \le \|v_0 - z_0\|_{\infty,\Omega} + \int_0^t K_\eta \|w(\cdot\,, s)\|_{\infty,\Omega} ds.$$

Therefore, since $\|w(\cdot\,,t)\|_{\infty,\Omega} = \max_{1\le i\le m} \|w_i(\cdot\,,t)\|_{\infty,\Omega}$, we have

$$\|w(\cdot\,,t)\|_\infty \le e^{K_\eta t}\|v_0 - z_0\|_\infty.$$

Consequently, because of (2.27) we have

$$\|w\|_{\infty,\Omega\times\mathbb{R}_+} \le \max\left\{e^{K_\eta(k+2)}\|v_0 - z_0\|_\infty,\ \tilde{\rho}_k(\|v_0 - z_0\|_\infty)\right\}. \tag{2.28}$$

Finally, since $\eta > 0$ is fixed, for any $\varepsilon \in (0, \eta)$ there exists $\delta > 0$ such that

$$\|v_0 - z_0\|_\infty < \delta \ \text{ implies } \ \|v - z_0\|_{\infty,\Omega\times\mathbb{R}_+} = \|w\|_{\infty,\Omega\times\mathbb{R}_+} < \varepsilon. \qquad \square$$

We point out that if z_0 is not a constant, we can modify the preceding arguments as follows. Suppose that $z_0 = w$ is a smooth function satisfying

$$-D\Delta w = f(w) \quad \text{on } \Omega, \tag{2.29a}$$

$$\partial w/\partial \mathbf{n} = 0 \quad \text{on } \partial\Omega \tag{2.29b}$$

In a manner similar to what was done above, the vector field may be truncated in a rectangular neighborhood containing $\{w(x) \mid x \in \Omega\}$. For $\eta > 0$, let $b_1(\eta, w)$ be an m-dimensional cube such that $w \in \text{int} b_1(\eta, w)$ with $\eta = \inf_{x\in\Omega} \operatorname{dist}(w(x), \partial b_1(\eta, w))$, and let $b_2(\eta, w)$ denote the m-cube concentric to $b_1(\eta, w)$ with twice the diameter. We mollify the characteristic function of $b_1(\eta, w)$ to produce a nonnegative function $\varphi_{\eta,w}$ such that

(i) $\varphi_{\eta,w} \in C^\infty(\mathbb{R}^m; [0, 1])$,
(ii) $\varphi_{\eta,w}(u) = 1$ if $u \in b_1(\eta, w)$, and
(iii) $\varphi_{\eta,w}(u) = 0$ if $u \in \mathbb{R}^m \setminus b_2(\eta, w)$,

and thus produce a corresponding truncated system (cf. (2.8) and (2.9)):

$$\begin{aligned} &\partial v/\partial t = D\Delta v + f[\eta, w](v) && \text{on } \Omega \times (0, \infty), \\ &\partial v/\partial \mathbf{n} = 0 && \text{on } \partial\Omega \times (0, \infty), \\ &v(\cdot\,, 0) = \varphi_{\eta,w}(u_0)u_0 && \text{on } \Omega. \end{aligned} \tag{2.30}$$

If $\mu = v - w$ and η is chosen such that $\eta > \|w\|_\infty$, we have

$$\begin{aligned} &\partial\mu_i/\partial t = d_i\Delta\mu_i + f_i[\eta, w](v) - f_i[\eta, w](w) && \text{on}\Omega \times (0, \infty), \\ &\partial\mu_i/\partial \mathbf{n} = 0 && \text{on}\partial\Omega \times (0, \infty), \\ &\mu_i(x, 0) = v_{0_i} - w_i && \text{on}\Omega. \end{aligned}$$

Then it is not difficult to establish an analogue of Lemma 2.4 and deduce that global solutions to (2.30) exist and that if they are sufficiently close to w, they satisfy (2.30). The following result concludes this section. Its proof is essentially a verbatim repetition of the one given for Theorem 2.6.

THEOREM 2.7. *Let $w \in M$ be a classical, spatially nonhomogeneous solution to the elliptic system* (2.29). *If $r > 0$ and w is a uniformly ∞–r stable steady state of* (2.30)*, then w is a uniformly stable steady-state solution of* (1.2). *Analogous results hold for uniformly ∞–r asymptotically stable solutions.*

We remark that an interesting reference pertaining to (2.29) is Matano [15].

3. *D*-diffusively convex Lyapunov functionals.

The most common tool for analyzing he local stability of equilibrium points for systems of ordinary differential equations of the form of (1.1) is the principle of linearized stability. If all the eigenvalues of the derivative of f at z_0 have negative real part, then z_0 is locally asymptotically stable. On the other hand, if any of the eigenvalues have positive real part, then the equilibrium point z_0 is unstable. These ideas carry over to the context of semilinear parabolic equations; see, e.g., [9]. In the case of nonhyperbolic equilibrium points, however, linearization methods do not apply.

Questions of nonlinear stability are frequently resolved by Lyapunov's direct method. Roughly speaking, a Lyapunov function V is a nonnegative functional which is defined and continuously differentiable in a neighborhood of a equilibrium point z_0 and is uniquely minimized in that neighborhood by z_0. If

$$\dot{V}(u) = \partial V(u) f(u) \leq 0 \tag{3.1}$$

in this neighborhood, then it follows that z_0 is a stable equilibrium point. Asymptotic stability can be deduced from conditions such as

$$\dot{V}(u) < -\alpha V(u) \tag{3.2}$$

for some $\alpha > 0$. In certain cases, a Lyapunov functional satisfying (3.1) in a neighborhood of an equilibrium point of a system of ordinary differential equations is useful in the context of the associated reaction-diffusion system. For this purpose, we introduce the notion of D-diffusively convex Lyapunov functionals for reaction-diffusion systems.

DEFINITION 3.1. *Let D be the matrix of diffusion coefficients for* (1.2) *and suppose that M is a forward invariant rectangle* (*possibly unbounded*) *for* (1.2). *If $z_0 \in M$ is an equilibrium point of f, we say that a nonnegative functional V is a D-diffusively convex Lyapunov functional around z_0 provided that the following conditions hold:*

(i) *There exists a $\xi > 0$ such that $V \in C^2(M \cap B_\xi(z_0); \mathbb{R}^+)$.*

(ii) *There exist constants $r > 0$ and $K > 0$ such that*

$$V(u) \geq K \sum_{i=1}^{m} |u_i - z_{0_i}|^r \quad \textit{for } u \in B_\xi(z_0) \cap M.$$

(iii) $V(z_0) = 0$.

(iv) *The matrix $D\partial^2 V(u)$ is positive semidefinite for $u \in B_\xi(z_0) \cap M$.* (*Here $\partial^2 V(u)$ is the Hessian matrix of V.*)

(v) $\partial V(u) f(u) \leq 0$ *for $u \in B_\xi(z_0) \cap M$.*

We remark that conditions (i)–(iii) and (v) are essentially those which define a Lyapunov functional for (1.1) around z_0 and that condition (iv) represents an additional strengthening of the concept. If the functional V is separable, i.e.,

$$V(u) = \sum_{i=1}^{n} V_i(u_i), \tag{3.3}$$

then we may ensure (iv) by assuming that $V_i''(u_i) \geq 0$. In general, however, convexity of V does not suffice for condition (iv). It is relatively straightforward to see that D-diffusively convex Lyapunov functionals guarantee the persistence of stability of equilibrium points. We have the following theorem.

THEOREM 3.2. *Let $z_0 \in M$ be an equilibrium point for the vector field f, where M is a forward invariant set for the semilinear parabolic system* (1.2). *If there exists a D-diffusively convex Lyapunov functional V for f around z_0, then z_0 is a stable steady state for* (1.2) *with respect to M. Moreover, if V also satisfies* (3.2), *then z_0 is asymptotically stable with respect to M.*

Proof. We choose $\eta > 0$ so that the cube $C_{2\eta}(z_0)$ is contained in $B_\xi(z_0)$, and we construct the truncated vector field $f[\eta, z_0]$ as in (2.8) and (2.9). If $v_0(x) \in C_{2\eta}(z_0) \cap M$ for $x \in \Omega$, it is immediately verified that

$$\partial V(v(x,t)) f(v(x,t)) = \partial V(v(x,t)) f[\eta, z_0](v(x,t)) \leq 0. \tag{3.4}$$

If we multiply the ith component of (2.9a) by $\partial V(v)/\partial v_i$, we obtain

$$(\partial V(v)/\partial v_i) \partial v_i/\partial t = d_i (\partial V(v)/\partial v_i) \Delta v_i + (\partial V(v)/\partial v_i) f_i[\eta, z_0](v). \tag{3.5}$$

If we integrate this expression on the space–time cylinder and sum the components, we observe that

$$\begin{aligned} \int_\Omega V(v(x,t))dx = &- \int_0^T \int_\Omega (\nabla v)^T D \partial^2 V(v) \nabla v dx dt \\ &+ \int_0^t \int_\Omega \partial V(v) f[\eta, z_0](v) dx + \int_\Omega V(v_0(x))dx. \end{aligned}$$

Hence by virtue of conditions (iv) and (v) in Definition 3.1, we have

$$\int_\Omega V(v(x,t))dx \leq \int_\Omega V(v_0(x))dx. \tag{3.6}$$

Using (3.6) and the coercivity of V, we get

$$\begin{aligned} K \left[\sum_{i=1}^{m} \|v_i(\cdot\,, t) - z_{0,i}\|_{r,\Omega} \right] &\leq \left[\int_\Omega V(v(x,t))dx \right]^{1/r} \\ &\leq \left[\int_\Omega V(v_0(x))dx \right]^{1/r} \\ &\leq \rho \left(\sum_{i=1}^{m} \|v_{0_i} - z_{0_i}\|_{\infty,\Omega} \right) \end{aligned} \tag{3.7}$$

for some continuous ρ with $\rho(0) = 0$ and $\rho(s) > 0$ for $s > 0$. This will ensure ∞–r stability, and from Theorem 2.6 we may conclude that z_0 is stable. Finally, in case

(3.2) holds, we take v_0 sufficiently close to z_0 to guarantee that our solution stays close to z_0 for all $t > 0$. Then we can obtain the estimate

$$\int_\Omega V(v(x,t))dx \leq e^{-\alpha t} \int_\Omega V(v_0(x))dx, \tag{3.8}$$

and from this follows the asymptotic stability assertion. □

In view of Theorem 2.7, one can naturally be lead to attempt to use D-diffusively convex Lyapunov functions to analyze the stability of spatially nonhomogeneous steady-state solutions. The following simple proposition squashes this endeavor for large classes of dynamical systems.

PROPOSITION 3.3. *Let M be a forward invariant set for* (1.2) *and let $V(v) = \sum_{i=1}^m V_i(v_i)$ be a nonnegative separable function which satisfies the defining hypotheses of Definition* 3.1, *except possibly* (ii) *and* (iii), *for all points of M. If $w = (w_1, \ldots, w_m)^T \in M$ is a solution to* (2.29), *then the following are true:*

(i) *If there exists $\alpha > 0$ such that $V_i''(v_i) > \alpha$ for all $v = (v_1, \ldots, v_m)^T \in M$, then $f(w) = 0$.*

(ii) *If $V(v) = \sum_{i=1}^m c_i v_i$, $\partial V(v) f(v) \leq 0$ and $M \subseteq \mathbb{R}_+^m$, then there exists a $k > 0$ such that $y(x) = \sum_{i=1}^m c_i d_i w_i(x) = k$ for all $x \in \Omega$, i.e., $w(x)$ belongs to a closed bounded subset of the hyperplane $\{v \mid \Sigma c_i d_i v_i = k\} \cap \mathbb{R}_+^m$.*

Proof. In the first case, we multiply the ith component of (2.29a) by $V_i'(w_i)$ to obtain

$$-d_i V_i'(w_i) \Delta w_i = V_i'(w_i) f_i(w). \tag{3.9}$$

If we sum these terms and integrate on Ω, we have

$$\sum_{i=1}^m d_i \int_\Omega V_i''(w_i) |\nabla w_i|^2 dx = \sum_{i=1}^n \int_\Omega V_i'(w_i) f_i(w) dx \leq 0. \tag{3.10}$$

Consequently,

$$\sum_{i=1}^m \alpha d_i \int_\Omega |\nabla w_i|^2 dx = 0, \tag{3.11}$$

and we may conclude that each w_i is a constant. Therefore, because $w = (w_1, \ldots, w_n)^T$ is a solution to (2.29), we must have $f_i(w) = 0$. If we follow the same train of reasoning for the second case, then we observe that $-\Delta(\Sigma c_i d_i w_i) \leq 0$. The fact that M is required to lie in $\mathbb{R}_+^m$ implies that $\Sigma d_i w_i \geq 0$, and hence we conclude from maximum principles that $\nabla(\Sigma c_i d_i w_i)$ vanishes and $\Sigma c_i d_i w_i(x) = k$ for some constant $k \geq 0$. Thus $w(x)$ lies in the hyperplane $\{v \mid \Sigma c_i d_i v_i = k\}$. The continuity of y implies that its range is closed and bounded. □

As a closing remark for this section, we point out that an additional treatment of Lyapunov theory in the context of reaction-diffusion systems can be found in [20].

4. Applications.

We begin by considering of the two-component system

$$\begin{aligned}
&\partial u/\partial t - d_1 \Delta u = -f(u,v) && \text{on } \Omega \times (0,\infty),\\
&\partial v/\partial t - d_2 \Delta v = f(u,v) && \text{on } \Omega \times (0,\infty),\\
&\partial u/\partial \mathbf{n} = \partial v/\partial \mathbf{n} = 0 && \text{on } \partial\Omega \times (0,\infty),\\
&u(\,\cdot\,,0) = u_0(\,\cdot\,), \qquad v(\,\cdot\,,0) = v_0(\,\cdot\,) && \text{on } \Omega,
\end{aligned} \tag{4.1}$$

where $f \in C^2(\mathbb{R}^2_+; \mathbb{R}_+)$ and $f(0,v) = 0$ for all $v \in \mathbb{R}_+$. Here we assume that the initial data u_0 and v_0 are continuous and nonnegative on $\overline{\Omega}$. It may be surprising that questions concerning the global existence of solutions to this system remain open. If the nonlinearity f is polynomially bounded, then it is known [10] that solutions to (4.1) exist for all time and remain uniformly bounded in the $L_\infty(\Omega)$ norm. Analogous results [8] have also have been obtained in the case where the nonlinearity is of the form

$$f(u,v) = u\varphi(v), \tag{4.2}$$

where φ need not be polynomially bounded but is required to grow less than exponentially, e.g., $\varphi(v) = e^{\sqrt{v}}$.

We are able to establish a simple result concerning the stability of the steady state $(0, \tilde{v})$ for (4.1).

PROPOSITION 4.1. *If $\tilde{v} \geq 0$, then the constant solution $(u,v) = (0,\tilde{v})$ is a stable equilibrium point for* (4.1) *with respect to $\mathbb{R}^2_+$.*

Proof. By assumption, $f(0,\tilde{v}) = 0$, and hence $(0,\tilde{v})$ is a steady-state solution of the system. In the case where $\tilde{v} = 0$, the result follows by noting that $\mathbb{R}^2_+$ is an invariant m-cube for the system, that ∞–1 stability follows from integrating each equation on the space–time cylinder and adding them to obtain the conservation law

$$\int_\Omega (u(x,t) + v(x,t))dx = \int_\Omega (u_0(x) + v_0(x))dx, \tag{4.3a}$$

and that $V = u + v$ defines a D-diffusively convex Lyapunov functional around $(0,0)$ with respect to $\mathbb{R}^2_+$. Now suppose that $\tilde{v} > 0$ and let $0 < \varepsilon < \tilde{v}$. Maximum principles demonstrate that solutions which initially lie in $M_\varepsilon = \{(u,v) \mid u \geq 0,\ v \geq \tilde{v} - \varepsilon\}$ remain so. The conservation law

$$\int_\Omega (u(x,t) + v(x,t) - (\tilde{v} - \varepsilon))dx = \int_\Omega (u_0(x) + v_0(x) - (\tilde{v} - \varepsilon))dx \tag{4.3b}$$

follows as before. Thus if the initial data are close to $(0, \tilde{v} - \varepsilon)$ in the L^∞ norm, then the solution remains close in the L^1 norm. Also, $V = u + v - (\tilde{v} - \varepsilon)$ defines a D-diffusively convex Lyapunov functional around $(0, \tilde{v} - \varepsilon)$ with respect to M_ε. Consequently, ∞–1 stability implies uniform stability with respect to M_ε. Therefore, it follows that solutions in $\mathbb{R}^m_+$ can be made to remain uniformly close to $(0, \tilde{v})$. □

We hope that we do not belabor the issue by pointing out that the system

$$\begin{aligned} \partial u/\partial t &= d_1 \Delta u - u e^{kv^\gamma}, \\ \partial v/\partial t &= d_2 \Delta v + u e^{kv^\gamma}, \end{aligned} \tag{4.4}$$

for example, with $\partial u/\partial \mathbf{n} = \partial v/\partial \mathbf{n} = 0$ on $\partial\Omega$ and any $\gamma \geq 1$ satisfies the hypotheses and hence admits $(0,\tilde{v})$ as a stable solution with respect to M as above whenever $\tilde{v} \geq 0$.

We now focus on a general class of diffusive Lotka–Volterra systems. Typically, Lotka–Volterra systems feature quadratic nonlinearities. They are intended to describe the species interaction among m-species ecological systems. Here we follow the development of Leung [13] and consider systems of the form

$$\begin{aligned} &\partial u/\partial t = D\Delta u + U(e + Pu) && \text{on } \Omega \times (0,\infty), \\ &\partial u/\partial \mathbf{n} = 0 && \text{on } \partial\Omega \times (0,\infty), \\ &u(\,\cdot\,, 0) = u_0(\,\cdot\,), && \text{on } \Omega, \end{aligned} \tag{4.5}$$

where $U = \operatorname{diag}\{u_1, \dots, u_m\}$, $e = (e_1, \dots, e_m)^T$ is a constant vector, and $P = (p_{ij})$ is an $m \times m$ matrix with constant entries. We assume that the following conditions are satisfied.

$(\mathrm{LV})_1$ There is a vector $q = (q_1, \dots, q_m)^T$, with each $q_i > 0$, that solves the linear system

$$e + Pq = 0. \tag{4.6}$$

$(\mathrm{LV})_2$ For each q satisfying (4.6) there is a diagonal matrix $A = \operatorname{diag}\{a_1, \dots, a_m\}$, with each $a_i > 0$, such that for all $w \in \mathbb{R}^m$,

$$(Aw)^T Pw = \sum_{i,j=1}^{m} a_i w_i p_{ij} w_j \leq 0. \tag{4.7}$$

Condition $(\mathrm{LV})_1$ guarantees the existence of a steady state with positive components. However, we have made no assumptions concerning the nonsingularity of the matrix P. Indeed, many Lotka–Volterra systems feature a multiplicity of positive steady states. The nonnegativity of the quadratic form (4.7) translates as weighted conservation of the interaction between the species of the system. Leung refers to this condition as admissibility.

The next lemma asserts that a well-known Lyapunov function for (4.5) provides a D-diffusively convex Lyapunov structure.

LEMMA 4.2. *There exists $\xi > 0$ such that the function V on $\mathbb{R}^m_+ \cap B_\xi(q)$ defined by*

$$V(v) = \sum_{i=1}^{m} V_i(v_i) = \sum_{i=1}^{m} \big(a_i(u_i - q_i) - a_i q_i \log(u_i/q_i)\big) \tag{4.8}$$

is D-diffusively convex on $\mathbb{R}^m_+ \cap B_\xi(q)$.

Proof. If $B_\xi(q)$ does not intersect the coordinate hyperplanes of $\mathbb{R}^m_+$, then it is clear that V is continuously differentiable and nonnegative on $B_\xi(q)$. Moreover, it is clear that $V(q) = 0$, and a careful analysis will reveal that $K > 0$, $\xi > 0$, and $r > 0$ may be chosen so that hypothesis (ii) of Definition 3.1 holds. We observe that if $u \in \mathbb{R}^m_+$, then

$$\partial V(u) f(u) = (A(u - q))^T P(u - q) = \sum_{i,j=1}^{m} a_i(u_i - q_i) P_{ij}(u_j - q_j) \leq 0. \tag{4.9}$$

The separability of V and the observation that $V_i''(v_i) = a_i q_i / v_i$ complete the proof. □

We immediately have the following result.

PROPOSITION 4.3. *If $(\mathrm{LV})_1$ and $(\mathrm{LV})_2$ are satisfied, then the steady-state solution $q = (q_1, \dots, q_m)^T$ is stable. Moreover, the semilinear elliptic system*

$$\begin{aligned} -D\Delta w &= W(e + Pw) && \text{on } \Omega, \\ \partial w/\partial \mathbf{n} &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{4.10}$$

where $W = \operatorname{diag}\{w_1, \dots, w_m\}$, has no spatially nonhomogeneous positive solutions.

Proof. Because $\mathbb{R}^m_+$ is an invariant m-cube for solutions to (4.5), Lemma 4.2 and Theorem 3.2 establish the first assertion. To establish the second assertion, we let M_1

be an m-cube which contains both $w \in \mathbb{R}^m_+$ and $q \in \mathbb{R}^m_+$ and does not intersect the coordinate hyperplanes of $\mathbb{R}^M_+$, and we let M_2 be a second m-cube which contains M_1 and also does not intersect the coordinate hyperplanes. Now let $\varphi \in C^\infty(\mathbb{R}^m; \mathbb{R}_+)$ be such that $\varphi(u) = 1$ if $u \in M_1$ and $\varphi(u) = 0$ for $u \in \mathbb{R}^m \setminus M_2$. From the application of part (i) of Proposition 3.3 to the truncated system

$$\begin{aligned} -D\Delta w &= \varphi(w) W(e + Pw) \quad \text{on } \Omega, \\ \partial w/\partial \mathbf{n} &= 0 \qquad\qquad\qquad\quad\; \text{on } \partial\Omega, \end{aligned} \tag{4.11}$$

the remaining assertion follows directly. □

We mention that for $n \geq 3$, the question of global existence for (4.5) is in general unresolved. For spatial dimension one, global existence and uniform boundedness for solutions may be established by applying results in [16], and for $n = 2$, we are at least assured the existence of long-time solutions; see [17]. We mention this to underscore the point that global well-posedness theory for reaction-diffusion systems remains incomplete.

Differential equations which describe the dispersion and reaction of m chemical species are generally of the form

$$\partial u/\partial t = D\Delta u + f(u), \tag{4.12}$$

where the ith component of the dependent variable $u = (u_1, \ldots, u_m)^T$ represents the concentration density of the ith chemical species. The vector field $f = (f_i)_{i=1}^m$ is assumed to be in each component a polynomial function of the components of u and is intended to model the chemical reaction kinetics. In his study of dissipative chemical reactions [7], Gröger introduced the following hypothesis.

(G) There exists a vector $e = (e_1, \ldots, e_m)^T$ with each $e_i > 0$ such that $f(e) = 0$ and

$$\sum_{i=1}^m f_i(u) \log(u_i/e_i) \leq 0.$$

Furthermore, the quantity $\sum_{i=1}^m f_i(u) \log(u_i/e_i)$ is known to have the physical interpretation of being a suitably scaled rate of chemical dissipation, and work on the mathematical theory of reaction networks [11] confirms that many nontrivial systems satisfy this hypothesis. If the chemical species are required to remain confined to a reaction vessel for all time, the appropriate boundary conditions are given by

$$\partial u/\partial \mathbf{n} = 0 \quad \text{on } \partial\Omega \times (0, \infty). \tag{4.13}$$

Finally, a condition of the form

$$f_i(u) \geq 0 \quad \text{for all } u \in \mathbb{R}^m_+ \quad \text{with } u_i = 0 \tag{4.14}$$

together with the maximum principle ensures that $\mathbb{R}^m_+$ is a forward invariant set for (4.12). We have the following proposition.

Proposition 4.4. *We consider* (4.12) *together with the boundary conditions* (4.13). *If all the conditions describing a dissipative chemical reaction outlined above hold, then the steady state $u = e$ is uniformly stable. Moreover, the elliptic system*

$$\begin{aligned} -D\Delta w &= f(w) \quad \textit{on } \Omega, \\ \partial w/\partial \mathbf{n} &= 0 \qquad\;\; \textit{on } \partial\Omega \end{aligned} \tag{4.15}$$

has no spatially inhomogeneous positive solutions.

Proof. We define

$$V(u) = \sum_{i=1}^{m} V_i(u) = \sum_{i=1}^{m} (u_i \log(u/e_i) - u_i + e_i) \tag{4.16}$$

and verify that all of the conditions of Definition 3.1 hold locally about e. Consequently, Theorem 3.2 implies that e is uniformly stable. An argument analogous to the one of Proposition 4.3 ensures the nonexistence of positive spatially inhomogeneous steady states. □

The comments concerning the global well-posedness and boundedness of solutions to Lotka–Volterra systems also apply to this class of dissipative chemical systems.

In addition to satisfying $f(0) = 0$ and a condition of the form (4.4), many reaction-diffusion systems satisfy a linear balancing condition of the following form.

(B) There exist positive constants c_i for $i =$ to m such that for all $n \in \mathbb{R}^m_+$,

$$\sum_{i=1}^{m} c_i f_i(u) = 0.$$

In this case, an obvious generalization of Proposition 4.1 dictates the stability of the zero solution.

5. Further generalizations and concluding remarks. Our results tend to support the general hypothesis that the addition of diffusion to systems of ordinary differential equations which have D-diffusively convex Lyapunov functions does not create exotic spatial or temporal phenomena which did not originally exist. If this is indeed the case, then the presence of diffusion in these systems is irrelevant to their long-term dynamics, and any spatial phenomena produced by diffusion must be of a transient nature.

We need not have limited our consideration to diffusion mechanisms of the form $D\Delta u$. We could have allowed operators of the form

$$\sum_{j,k=1}^{n} \frac{\partial}{\partial x_k} \left(d^i_{jk}(x,t) \frac{\partial u_i}{\partial x_j} \right)$$

in each component. In this case, it is necessary to assume uniformly strong ellipticity along with smoothness conditions on coefficients and some conditions on the derivatives of the coefficients. In general, the arguments could become quite technical but should be tractable. We leave the details to the interested reader. Numerical experiments [4] with two-component systems which model exothermic chemical reactions indicate that quasi-linear diffusivities do have an effect on the intermediate dynamics of the systems.

The necessity that our forward invariant set M be an m-cube described by (2.4) is purely a consequence of assuming distinct diffusion coefficients and in no way actually enters into the preceding analysis. Other types of geometries can arise in situations where some of the diffusion coefficients are equal. As a simple example, consider a three-component model of the form

$$\begin{aligned} \partial u/\partial t - \Delta u &= -\alpha_1 f(u,v,w) && \text{on } \Omega \times (0,\infty), \\ \partial v/\partial t - \Delta v &= -\alpha_2 f(u,v,w) && \text{on } \Omega \times (0,\infty), \\ \partial w/\partial t - d\Delta w &= f(u,v,w) && \text{on } \Omega \times (0,\infty), \\ \partial u/\partial \mathbf{n} = \partial v/\partial \mathbf{n} &= \partial w/\partial \mathbf{n} = 0 && \text{on } \partial\Omega \times (0,\infty), \\ u(\,\cdot\,,0) &= u_0, \quad v(\,\cdot\,,0) = v_0, \quad w(\,\cdot\,,0) = w_0 && \text{on } \Omega, \end{aligned} \tag{5.1}$$

where $f \in C^2(\mathbb{R}^3_+;\mathbb{R}_+)$; $f(0,v,w) = f(u,0,w) = 0$ for all $u, v, w \in \mathbb{R}_+$; $\alpha_1, \alpha_2, d > 0$; and the initial data u_0, v_0, and w_0 are continuous and nonnegative on $\overline{\Omega}$. By the maximum principle, it follows that $\min_\Omega\{\alpha_1 v_0 - \alpha_2 u_0\} \le \alpha_1 v - \alpha_2 u \le \max_\Omega\{\alpha_1 v_0 - \alpha_2 u_0\}$ and $w \ge \min_\Omega w_0$. Consequently, if $z_1, z_3 \ge 0$, then the set

$$M_1 = \{(u,v,w) \mid \alpha_1 v - \alpha_2 u \le -\alpha_2 z_1, \quad v \ge 0, \quad w \ge z_3\}$$

is a forward invariant set for (5.1), and if $z_2, z_3 \ge 0$, then the set

$$M_2 = \{(u,v,w) \mid u \ge 0, \quad \alpha_1 v - \alpha_2 u \ge \alpha_1 z_2, \quad w \ge z_3\}$$

is a forward invariant set for (5.1). Now in a manner similar to the proof of Proposition 4.1, one can show that any point $(z_1, 0, z_3)$ with $z_1, z_3 \ge 0$ is stable with respect to M_1 and any point $(0, z_2, z_3)$ with $z_2, z_3 \ge 0$ is stable with respect to M_2. One can then continue to argue as in the proof of Proposition 4.1 that such equilibrium points are stable with respect to $\mathbb{R}^3_+$.

REFERENCES

[1] N. Alikakos, *An application of the invariance principle to reaction-diffusion equations*, J. Differential Equations, 33 (1979), pp. 201–225.

[2] H. Amann, *Existence and regularity for semilinear parabolic evolution equations,* Ann. Scoula Norm. Sup. Pisa Cl. Sci. (4), IX (1984), pp. 593–676.

[3] J. F. G. Auchmuty, *Qualitative effects of diffusion in chemical systems*, Lectures Math. Life Sci., 10 (1978), pp. 49–99.

[4] W. Fitzgibbon and C. Martin, *The longtime behavior of solutions to a quasilinear combustion model,* J. Nonlinear Anal., 19 (1992), pp. 947–961.

[5] W. Fitzgibbon, J. Morgan, and R. Sanders, *Global existence and boundedness for a class of inhomogeneous parabolic equations,* J. Nonlinear Anal., 19 (1992), pp. 885–899.

[6] W. Fitzgibbon, J. Morgan, and S. Waggoner, *Weakly coupled semilinear parabolic evolution systems,* Ann. Mat. Pura Appl. (4), CLXI (1992), pp. 213–229.

[7] K. Gröger, *On the existence of steady-states of certain reaction-diffusion systems,* Arch. Rational Mech. Anal., 92 (1986), pp. 297–306.

[8] A. Haraux and A. Youkana, *On a result of K. Masuda concerning reaction-diffusion equations,* Tôhoku Math. J., 40 (1988), pp. 158–183.

[9] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[10] S. Hollis, R. Martin, and M. Pierre, *Global existence and boundedness in reaction-diffusion systems,* SIAM J. Math. Anal., 18 (1987), pp. 744–761.

[11] F. Horn and R. Jackson, *General mass action kinetics,* Arch. Rational Mech. Anal., 47 (1972), pp. 81–116.

[12] O. Ladyženskaja, V. Solonnikov, and N. Uralćeva, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl., Vol. 23, AMS, Providence, RI, 1968.

[13] A. Leung, *Systems of Nonlinear Partial Differential Equations*, Kluwer Academic Publishers, Boston, 1989.

[14] R. Martin and M. Pierre, *Nonlinear reaction diffusion systems,* in *Nonlinear Equations in the Applied Sciences*, W. F. Ames and C. Rogers, eds., Academic Press, New York, 1990.

[15] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations,* Publ. Res. Inst. Math. Sci., 15, (1979), pp. 401–454.

[16] J. Morgan, *Boundedness and decay results for reaction-diffusion systems,* SIAM J. Math. Anal., 21 (1990), pp. 1172–1181.

[17] J. Morgan, *Global existence for a class of quasilinear reaction-diffusion systems,* preprint.

[18] J. Murray, *Mathematical Biology,* Springer-Verlag, Berlin, 1989.

[19] M. Pierre and D. Schmidt, *Blowup in reaction-diffusion systems with dissipation of mass,* SIAM J. Math. Anal., 28 (1997), pp. 259–269.

[20] R. Redheffer, R. Redlinger, and W. Walter, *A theorem of LaSalle–Lyapunov type for parabolic systems,* SIAM J. Math. Anal., 19 (1988), pp. 121–132.

[21] J. Smoller, *Shock Waves and Reaction-Diffusion Equations,* Springer-Verlag, Berlin, 1984.

[22] H. Stewart, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions,* Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 611–632, May 1997

AN INVERSE PROBLEM FOR THE HYDRAULIC PROPERTIES OF POROUS MEDIA*

PAUL DuCHATEAU†

Abstract. An inverse problem is formulated to determine the two coefficients in the pressure head formulation of the porous flow equation from a simple hydraulic experiment. Integral identities are derived which relate changes in the coefficients to changes in measured outputs. These identities are used to precisely define the sense in which the experimental data are able to distinguish between different porous media. It is also shown that the mapping associating input coefficient values to output data values is explicitly invertible and that there is a related output least squares problem whose solution is the solution of the inverse problem.

Key words. inverse problems, parabolic equations, structural identification

AMS subject classifications. 35R30, 35Q80, 35K60

PII. S0036141095285673

Introduction. Flow in an unsaturated porous medium can be modeled by nonlinear partial differential equations in which the coefficients in the equation characterize the hydraulic properties of the medium. Treatment of such equations is simplified by assuming that the coefficients are functions of the unknown dependent variable only. This is equivalent to supposing the medium is homogeneous and isotropic, and in such cases it is feasible to seek coefficients which are characteristic of a specific porous medium [2, 5, 7, 10].

Since direct experimental measurement of hydraulic properties of porous media is often inconvenient, attempts have been made to obtain the properties indirectly by formulating and solving a suitable inverse problem. The goal of the indirect approach is to replace a difficult physical experiment by an inverse problem for which the input data are easy to measure and whose solution leads to the hydraulic properties of the medium. The requirement that the data be easy to measure suggests that the data should be dynamic rather than steady state and should be measured on the boundary of the physical domain [5, 12].

An obstacle to considering inverse problems involving nonlinear partial differential equations is the lack of an explicit solution for the so-called direct problem. One approach, referred to in the literature as the method of output least squares, seeks to avoid this obstacle by formulating the inverse problem as an optimization problem seeking coefficients which produce a solution for the differential equation that best matches some experimentally measured output [4, 11]. The appeal of the output least squares approach lies in the well-developed theory for dealing with optimization problems. On the other hand, it is usually not evident that the solution to the optimization problem solves the original inverse problem. In particular, the error functional may be based on data which do not uniquely determine the unknown coefficients. In [6, 8], examples are constructed which illustrate the difficulties with this approach.

This paper formulates an inverse problem to determine simultaneously the two coefficients in the pressure head formulation of the porous flow equation from a simple

*Received by the editors May 4, 1995; accepted for publication (in revised form) January 18, 1996.

http://www.siam.org/journals/sima/28-3/28567.html

†Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (pauld@math.colostate.edu).

hydraulic experiment. These coefficients, water capacity and hydraulic conductivity, characterize the hydraulic properties of a porous medium. Integral identities are derived which relate changes in the coefficients to changes in measured outputs. Using these identities it is possible to precisely define the sense in which the experimental data are able to distinguish between porous media. In addition it can be shown that a coefficient pair that optimizes an output least squares functional whose definition is based on the measured outputs of this inverse problem must necessarily also solve the inverse problem. Finally, the mapping associating coefficient values to output data values is shown to be explicitly invertible in a suitable class of coefficient pairs. More precisely, it is shown that if the measured data satisfy certain necessary conditions then it is possible to construct coefficients which approximately reproduce this output when used in the model equations. If the output is known to have been generated by a suitable coefficient pair, then it can be shown that the constructed coefficient pair approximates the actual coefficients.

Two experiments and the associated inverse problems are described here. In the first experiment, termed the phase one experiment, an initially saturated vertical column of soil is allowed to drain to equilibrium under gravity. Solution of the associated phase one inverse problem determines the hydraulic functions for the soil in the column over a portion of their domain. A second experiment, labeled phase two, leads to an inverse problem whose solution extends the portion of the domain over which the hydraulic functions are determined. These experiments are probably not unique but illustrate what seems to be an important feature for successful identification of unknown ingredients in partial differential equations where the ingredients are functions of the state variable only. The identification is considerably simplified if it is possible to construct an experiment where the boundary value of the state variable varies monotonically with time. In the two experiments described here, the boundary measurement of the pressure head $p(t) = h(0,t)$ is forced to behave monotonically; gravity is the driving force in phase one, and in the phase-two experiment it is the applied suction.

The organization of this paper is as follows. Section 1 establishes several essential facts about the solution of the so-called direct problem associated with the phase-one experiment and derives key integral identities for analyzing the inverse problem. Uniqueness and solvability results for the phase-one inverse problem are presented in section 2. Sections 3 and 4 repeat these procedures for the phase-two experiment.

1. The phase-one direct problem.

Consider a vertical soil column which is totally saturated and then allowed to drain under gravity. If there is no flow across the top end of the column and if the bottom end of the column is at the water table, then the capillary pressure head $h(z,t)$ can be shown to satisfy

$$
\begin{aligned}
&C(h)\partial_t h(z,t) = \partial_z(K(h)(\partial_z h(z,t) - 1)) && \text{for } 0 < z < L, \quad 0 < t < T,\\
&h(z,0) = 0 && \text{for } 0 < z < L,\\
&\partial_z h(0,t) - 1 = 0, \qquad h(L,t) = 0 && \text{for } 0 < t < T.
\end{aligned}
\tag{1.1}
$$

Here C and K denote the water capacity and hydraulic conductivity, respectively. The column is assumed to be of length L with $z = 0$ at the top of the column and $z = L$ at the bottom. For notational convenience, let $Q_T = \{(z,t)\colon\ 0 < z < L, 0 < t < T\}$.

Problem (1.1) will be called the phase-one direct problem. For suitable coefficients C and K this direct problem has a unique smooth solution [4, 9]. This solution is initially zero at each point and tends toward a steady state equal to the linear function $h(z) = z - L$, $0 < z < L$, as t tends to infinity. In point of fact, there is no finite time T at which $h(0, T)$ actually equals $-L$ but for every small positive ε there exists a finite time $T = T(\varepsilon)$ such that $h(0, T) = -L+\varepsilon$. Then as (z, t) ranges over Q_T, $h(z, t)$ takes its values in the interval $[-L + \varepsilon, 0]$ rather than in $(-L, 0)$. Recognizing this, the parameter T will nevertheless be assumed here to denote a fixed, large, positive number; the head values $h(z, t)$ will vary between zero and $-L$ during this phase of the experiment; and the ε will be omitted from subsequent discussions.

Coefficients C and K are said to be *admissible* if they satisfy

$$(1.2)\qquad \begin{aligned} &\text{(i)} \quad C\varepsilon\, \mathbb{C}(-\infty, 0] \quad \text{and} \quad 0 < c_0 \le C(h) \le c_1 \quad \text{for } h < 0,\\ &\text{(ii)} \quad K\varepsilon \text{ Piecewise}-\mathbb{C}^1(-\infty, 0] \quad \text{and} \quad 0 < k_0 \le K(h) \le k_1 \quad \text{for } h < 0. \end{aligned}$$

For each pair of admissible coefficients (C, K), the direct problem (1.1) has a unique solution $h(z, t)$ whose dependence on the coefficients will be indicated by the notation $h = \Psi_1[C, K]$. In a physical experiment in which a vertical column drains under gravity as described, it is relatively easy to measure the pressure head at the top of the column and to measure the flux or outflow at the bottom of the column. The draining of the column may be simulated by solving (1.1) for $h = \Psi_1[C, K]$, and the measured data then correspond to the computed functions

$$(1.3)\qquad p(t) = h(0, t) \quad \text{and} \quad q(t) = K(h(L, t))(\partial_z h(L, t) - 1) \quad \text{for } 0 < t < T.$$

The dependence of the outputs $p(t)$ and $q(t)$ on the coefficients C and K will be indicated by the notation $(p, q) = \Gamma \cdot \Psi_1[C, K]$. The notations $p = \Gamma_0 \cdot \Psi_1[C, K]$ and $q = \Gamma_L \cdot \Psi_1[C, K]$ indicate the association between the individual outputs and the coefficient pairs. This association defines the *coefficient-to-data mapping* for the phase-one experiment. Since the functions (p, q) are viewed as system outputs it is reasonable to expect that their properties are determined by, and must be deduced from, the equation and properties of the coefficients. These properties are of some interest in their own right, and they are essential to the analysis of the inverse problem described in subsequent sections of the paper. This first lemma characterizes the behavior of the output function $q(t)$ in the phase-one experiment.

LEMMA 1.1. *For admissible coefficients C and K, the output $q(t) = \Gamma_L \cdot \Psi_1[C, K]$ satisfies*

$$(1.4)\qquad q\varepsilon\mathbb{C}[0, T), \qquad q(0) = -K(0), \quad \text{and} \quad q(t) < 0 \quad \text{for } 0 < t < T.$$

Proof. The smoothness properties of the solution imply that $q(t)$ is continuous on $[0, T]$, and it then follows from the initial and boundary conditions that $q(0) = -K(0) \le -k_0$.

For $h = \Psi_1[C, K]$ and an arbitrary smooth function $\varphi(z, t)$,

$$(1.5)\qquad \iint_{Q_T} [\partial_t a(h(z, t)) - \partial_z(K(h)(\partial_z h(z, t) - 1))]\partial_z \varphi \, dz\, dt = 0,$$

where

$$a(h(z, t)) = \int_0^{h(z,t)} C(s)\, ds.$$

Then $\partial_t a(h(z,t)) = C(h)\partial_t h(z,t)$ and $a(h(z,0)) = 0$ for $h = \Psi_1[C,K]$. Integration by parts shows

$$\begin{aligned}\iint_{Q_T} \partial_t a(h(z,t))\partial_z\varphi\,dz\,dt &= \int_0^L a(h)\partial_z\varphi\Big|_{t=0}^{t=T} dz - \iint_{Q_T} a(h)\partial_{tz}\varphi\,dz\,dt\\ &= \iint_{Q_T} \partial_z hC(h)\partial_t\varphi\,dz\,dt - \int_0^T a(h)\partial_t\varphi\Big|_{z=0}^{z=L} dt\\ &\quad + \int_0^L a(h)\partial_z\varphi\Big|_{t=0}^{t=T} dz\end{aligned}$$

and

$$\begin{aligned}&\iint_{Q_T} [\partial_z(K(h)(\partial_z h(z,t)-1))]\partial_z\varphi\,dz\,dt\\ &\quad = \int_0^T (K(h)(\partial_z h(z,t)-1))\partial_z\varphi\Big|_{z=0}^{z=L} dt\\ &\qquad - \iint_{Q_T} (K(h)(\partial_z h(z,t)-1))\partial_{zz}\varphi\,dz\,dt.\end{aligned}$$

Then

$$\begin{aligned}&\iint_{Q_T} [(\partial_z h-1)(C(h)\partial_t\varphi + K(h)\partial_{zz}\varphi) + C(h)\partial_t\varphi]\,dz\,dt\\ &\quad = \int_0^T a(h)\partial_t\varphi + K(h)(\partial_z h-1)\partial_z\varphi\Big|_{z=0}^{z=L} dt - \int_0^L a(h)\partial_z\varphi\Big|_0^T dz.\end{aligned} \tag{1.6}$$

Now suppose $\varphi(z,t)$ solves the adjoint problem

$$\begin{aligned} C(h)\partial_t\varphi(z,t) + K(h)\partial_{zz}\varphi(z,t) &= 0 && \text{in } Q_T,\\ \varphi(z,T) &= 0, && 0 < z < L,\\ \varphi(0,t) &= 0, \qquad \partial_z\varphi(L,t) = \vartheta(t), && 0 < t < T\end{aligned}$$

for arbitrary smooth boundary input $\vartheta(t)$. Then

$$\begin{aligned} a(h(z,0)) &= 0 \quad\text{and}\quad a(h(L,t)) = 0,\\ \partial_t\varphi(0,t) &= 0 \quad\text{and}\quad \partial_z\varphi(z,T) = 0,\\ \partial_z h(0,t) - 1 &= 0,\end{aligned}$$

and it follows that (1.6) reduces to the following simple integral identity:

$$\int_0^T q(t)\vartheta(t)\,dt = \iint_{Q_T} C(h)\partial_t\varphi(z,t)\,dz\,dt.$$

Now choose the boundary input $\vartheta(t)$ in the adjoint problem such that $\vartheta(T) = 0$ and $\vartheta(t)$ is sufficiently large and positive for $0 < t < T$ that $\partial_{zz}\varphi(z,t) > 0$ on Q_T. Then $\partial_t\varphi(z,t) < 0$ on Q_T, and since $\vartheta(t)$ is otherwise arbitrary and $C(h(z,t)) \geq c_0 > 0$, it follows from the integral identity that $q(t) < 0$ almost everywhere on $[0,T]$.

Certain properties of the solution of the direct problem are needed if the integral identity arguments are to be used to analyze the inverse problem. The next lemma establishes one of these properties.

LEMMA 1.2. *For admissible coefficients $C(h)$ and $K(h)$, let $h = \Psi_1[C, K]$. Then $\partial_z h(z,t) - 1 < 0$ almost everywhere on Q_T.*

Proof. For $h = \Psi_1[C, K]$ and an arbitrary smooth function $\varphi(z,t)$, the basic integral identity (1.6) is valid. Suppose now that $\varphi(z,t)$ solves the adjoint problem

$$\begin{aligned} C(h)\partial_t\varphi(z,t) + K(h)\partial_{zz}\varphi(z,t) &= F(z,t) && \text{in } Q_T,\\ \varphi(z,T) &= 0, && 0 < z < L,\\ \varphi(0,t) = 0, \qquad \varphi(L,t) &= 0, && 0 < t < T. \end{aligned} \tag{1.7}$$

Since

$$a(h(z,0)) = 0, \quad \text{and} \quad \partial_z h(0,t) - 1 = 0$$

and

$$\partial_t\varphi(0,t) = 0, \qquad \partial_z\varphi(z,T) = 0, \quad \text{and} \quad \partial_t\varphi(L,t) = 0,$$

it follows that (1.6) reduces to

$$\iint_{Q_T} (\partial_z h - 1)F(z,t)\,dz\,dt = -\iint_{Q_T} C(h)\partial_t\varphi\,dz\,dt + \int_0^T q(t)\partial_z\varphi(L,t)\,dt. \tag{1.8}$$

If the function $F(z,t)$ appearing in the adjoint equation is nonnegative in Q_T, then the maximum principle asserts that the solution $\varphi(z,t)$ of (1.7) satisfies $\varphi(z,t) < 0$ in Q_T. If, in addition, at each z in $(0,L)$, the function $F(z,t)$ is assumed to increase so rapidly with respect to t that one has $\partial_{zz}\varphi(z,t) < 0$ in Q_T, then it is clear from the adjoint equation that $\partial_t\varphi(z,t) > 0$ in Q_T. Finally, $\varphi < 0$ in Q_T implies that $\partial_z\varphi(L,t) > 0$ for $0 < t < T$. Since $q(t)$ is already known to be negative by Lemma 1.1, it follows that the right side of (1.8) is strictly negative. $F(z,t)$ is nonnegative and increasing with t but is otherwise arbitrary; hence it follows from (1.8) that $\partial_z h(z,t) - 1 < 0$ almost everywhere in Q_T.

An additional property of the solution to the direct problem, also essential to the analysis of the inverse problem, is asserted in the following lemma.

LEMMA 1.3. *For admissible coefficients C and K, let $h = \Psi_1[C, K]$. Then $\partial_t h(z,t)$ is negative almost everywhere in Q_T.*

Proof. For $h = \Psi_1[C, K]$ and an arbitrary smooth function $\varphi(z,t)$,

$$\iint_{Q_T} [C(h)\partial_t h(z,t) - \partial_z(\partial_z b(h(z,t)) - K(h))]\partial_t\varphi\,dz\,dt = 0, \tag{1.9}$$

where

$$b(h(z,t)) = \int_0^{h(z,t)} K(s)\,ds.$$

Then integration by parts shows that

$$\begin{aligned} \iint_{Q_T} \partial_{zz} b(h(z,t))\partial_t\varphi\,dz\,dt &= \int_0^T \Big[\partial_z b(h)\partial_t\varphi + \partial_t b(h)\partial_z\varphi\Big]_{z=0}^{z=L} dt \\ &- \int_0^L \partial_z b(h)\partial_z\varphi\Big|_{t=0}^{t=T} dz - \iint_{Q_T} \partial_t b(h(z,t))\partial_{zz}\varphi\,dz\,dt \end{aligned}$$

and

$$\iint_{Q_T} \partial_z K(h(z,t))\partial_t\varphi\, dz\, dt = \int_0^T K(h)\partial_t\varphi\Big|_{z=0}^{z=L} dt - \int_0^L K(h)\partial_z\varphi\Big|_{t=0}^{t=T} dt + \iint_{Q_T} \partial_t K(h(z,t))\partial_z\varphi\, dz\, dt.$$

Then

$$\begin{aligned}&\iint_{Q_T} \partial_t h[C(h)\partial_t\varphi + K(h)\partial_{zz}\varphi + K'(h)\partial_z\varphi]\, dz\, dt\\ &= -\int_0^L K(h)(\partial_z h - 1)\partial_z\varphi\Big|_{t=0}^{t=T} dz \\ &\quad + \int_0^T \Big[(\partial_z b(h) - K(h))\partial_t\varphi + \partial_t b(h)\partial_z\varphi\Big]_{z=0}^{z=L} dt.\end{aligned} \tag{1.10}$$

If $\varphi(z,t)$ solves

$$\begin{aligned} C(h)\partial_t\varphi + K(h)\partial_{zz}\varphi + K'(h)\partial_z\varphi = F(z,t) \qquad & \text{in } Q_T,\\ \varphi(z,T) = 0, \qquad & 0 < z < L,\\ \partial_z\varphi(0,t) = 0, \qquad \partial_z\varphi(L,t) = 0, \qquad & 0 < t < T,\end{aligned} \tag{1.11}$$

then

$$\int_0^T \Big[(\partial_z b(h) - K(h))\partial_t\varphi + \partial_t b(h)\partial_z\varphi\Big]_{z=0}^{z=L} dt = \int_0^T q(t)\partial_t\varphi(L,t)\, dt$$

since $\partial_z b(h) - K(h) = K(h)(\partial_z h - 1)$ equals 0 and $q(t)$ at $z = 0$, L, respectively, and $\partial_t b(h(L,t)) = K(0)\partial_t h(L,t) = 0$. Also,

$$\int_0^L K(h)(\partial_z h - 1)\partial_z\varphi\Big|_{t=0}^{t=T} dz = -K(0)\int_0^L \partial_z\varphi(z,0)\, dz = K(0)(\varphi(0,0) - \varphi(L,0))$$

since $\partial_z\varphi(z,T) = 0$ and $\partial_z h(z,0) - 1 = -1$. Then (1.10) reduces to

$$\iint_{Q_T} \partial_t h(z,t)F(z,t)\, dz\, dt = \int_0^T q(t)\partial_t\varphi(L,t)\, dt + K(0)(\varphi(0,0) - \varphi(L,0)). \tag{1.12}$$

If $F(z,t)$ is nonnegative in Q_T, then the maximum principle applied to (1.11) asserts that $\varphi(z,t)$ can have no maximum on the interior of Q_T. Moreover, since $\partial_z\varphi = 0$ at $z = 0$, L, it follows [1, p. 261, 3] that $\varphi(z,t)$ cannot achieve a maximum at either of the boundary points $z = 0$, L. Then the maximum value of $\varphi(z,t)$ must occur at $t = T$, where $\varphi(z,T)$ is known to vanish for $0 < z < L$. It follows that $\varphi(z,t) < 0$ in Q_T. In particular, $\varphi(L,t) < 0$ for $0 < t < T$ and $\varphi(L,T) = 0$. The nonnegative, but otherwise arbitrary, function $F(z,t)$ may be chosen to grow so rapidly with respect to t near $z = L$ that $\partial_t\varphi(L,t) > 0$ for $0 < t < T$. In Lemma 1.1, $q(t)$ has been shown to be negative, so the first integral on the right side of (1.12) is strictly negative. Since F is nonnegative and increasing with t but is otherwise arbitrary, the values of F can be adjusted to cause $\varphi(0,0)-\varphi(L,0)\geq 0$. Then the right side of (1.12) is negative, and it follows that unless $\partial_t h(z,t)<0$ almost everywhere in Q_T it is possible to further adjust the nonnegative function F so that a contradiction of (1.12) is obtained.

The final lemma relating to properties of the solution to the direct problem characterizes the behavior of the output function $p(t)=h(0,t)$ during the phase-one experiment.

LEMMA 1.4. *For admissible coefficients C and K let $h(z,t)=\Psi_1[C,K]$. Then for each τ, $0<\tau\leq T$,*

$$z-L<h(z,t)<0 \quad \textit{and} \quad h(0,\tau)<h(z,t)<0 \quad \textit{for } 0<z<L, \quad 0<t<\tau.$$

In addition, $p(t)=h(0,t)=\Gamma_0\cdot\Psi_1[C,K]$ satisfies

$$p\varepsilon\mathbb{C}^1[0,T], \qquad p(0)=0, \quad \text{and} \quad p'(t)<0 \quad \text{for } 0<t<T. \tag{1.13}$$

Proof. For $h(z,t)=\Psi_1[C,K]$, let $u(z,t)=:h(z,t)-z+L$. Then for any τ, $0<\tau\leq T$, $u(z,t)$ solves

$$\begin{aligned}
C(h)\partial_t u &= \partial_z(K(h)\partial_z u) && \text{in } Q_\tau,\\
u(z,0) &= L-z, && 0<z<L,\\
\partial_z u(0,t) &= 0, \qquad u(L,t)=0, && 0<t<\tau.
\end{aligned}$$

The maximum–minimum principle implies that $0<u(z,t)<L-z$ in Q_τ for each τ in $(0,T]$; hence $z-L<h(z,t)<0$ in Q_T. Evidently, the maximum over the parabolic boundary for the function $h(z,t)$ occurs at $z=0$, where $\partial_z h(0,t)=1$, and this leads to $h(0,\tau)<h(z,t)<0$ for (z,t) in Q_τ, $0<\tau\leq T$.

To prove (1.13) choose the test function $\varphi(z,t)$ in the identity (1.10) to solve the following adjoint problem:

$$\begin{aligned}
C(h)\partial_t\varphi+K(h)\partial_{zz}\varphi+K'(h)\partial_z\varphi &= 0 && \text{in } Q_T,\\
\varphi(z,T) &= 0, && 0<z<L,\\
\partial_z\varphi(0,t) &= \vartheta(t), \qquad \varphi(L,t)=0, && 0<t<T.
\end{aligned}$$

Then (1.10) reduces to the identity

$$K(0)\int_0^L \partial_z\varphi(z,0)\,dz = -\int_0^T K(h(0,t))\partial_t h(0,t)\vartheta(t)\,dt;$$

i.e.,

$$K(0)\varphi(0,0)=\int_0^T K(h(0,t))p'(t)\vartheta(t)\,dt.$$

Choose the input $\vartheta(t)$ in the adjoint problem such that $\vartheta(T)=0$ and $\vartheta(t)<0$ for t in $[0,T)$. Then it follows from the extended maximum principle [1] that the maximum value of $\varphi(z,t)$ on the parabolic boundary of Q_T occurs at $z=0$. Evidently, $\varphi(0,t)$ is positive if $\vartheta(t)$ is negative for t in $[0,T)$; hence $\varphi(0,0)$ is positive and $K(0)\varphi(0,0)$ is positive as well. Since $\vartheta(t)$ is negative but is otherwise arbitrary, it follows from the last identity that $p'(t)$ must be negative for $0<t<T$.

Note that conditions (1.4) and (1.13) are necessary conditions for (p,q) to be a data pair generated by admissible coefficients as described by (1.3). A function pair (p,q) will be said to be an *admissible data pair* if they satisfy conditions (1.13) and

(1.4), respectively. The following theorem contains the integral identity on which the analysis of the inverse problem rests.

THEOREM 1.5. *For admissible coefficients C_j and K_j, $j = 1, 2$, let $h_j(z, t) = \Psi_1[C_j, K_j]$. For arbitrary smooth functions $\vartheta_0(t)$, $\vartheta_1(t)$, let the notation $\varphi = \Psi^*[\vartheta_0, \vartheta_1]$ indicate the solution of the following adjoint initial boundary value problem:*

$$\begin{aligned}&\alpha(z,t)\partial_t\varphi(z,t) + \beta(z,t)\partial_{zz}\varphi + \gamma(z,t)\partial_z\varphi = 0, && 0 < z < L, \quad 0 < t < \tau,\\ &\varphi(z,\tau) = 0, && 0 < z < L,\\ &\beta(0,t)\partial_z\varphi(0,t) = \vartheta_0(t), \qquad \varphi(L,t) = \vartheta_1(t), && 0 < t < \tau,\end{aligned} \tag{1.14}$$

where the coefficients α, β, γ are given by

$$\beta(z,t) = \int_0^1 K_1(h_2(z,t) + s(h_1(z,t) - h_2(z,t)))\,ds, \tag{1.15}$$

$$\alpha(z,t) = \int_0^1 C_1(h_2(z,t) + s(h_1(z,t) - h_2(z,t)))\,ds, \tag{1.16}$$

$$\gamma(z,t) = \int_0^1 K_1'(h_2(z,t) + s(h_1(z,t) - h_2(z,t)))\,ds. \tag{1.17}$$

If $(p_j, q_j) = \Gamma \cdot \Psi_1[C_j, K_j], j = 1, 2$, then increments in the inputs of $\Delta C = C_1 - C_2$, $\Delta K = K_1 - K_2$, lead to increments in output of $\Delta q = q_1 - q_2$, $\Delta p = p_1 - p_2$, and for any τ, $0 < \tau \le T$, the input and output increments are related by

$$\begin{aligned}&\int_0^\tau [\Delta q(t)\vartheta_1(t) + \Delta p(t)\vartheta_0(t)]\,dt\\ &\quad = \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z\varphi + \Delta C(h_2)\varphi\partial_t h_2]\,dz\,dt.\end{aligned} \tag{1.18}$$

Proof. For admissible coefficients C_j and K_j, let $h_j(z,t) = \Psi_1[C_j, K_j]$, $j = 1, 2$. Then for an arbitrary smooth function $\varphi(z,t)$ and any $\tau, 0 < \tau \le T$,

$$\begin{aligned}&\iint_{Q_\tau} \varphi[\partial_t(a_1(h_1) - a_1(h_2)) - \partial_z(\partial_z(b_1(h_1) - b_1(h_2)) - K_1(h_1) + K_1(h_2))]\,dz\,dt\\ &\quad = -\iint_{Q_\tau} \varphi[\partial_t(a_1(h_2) - a_2(h_2)) - \partial_z(\partial_z(b_1(h_2) - b_2(h_2)) - K_1(h_2) + K_2(h_2))]\,dz\,dt.\end{aligned}$$

Integration by parts leads to

$$\begin{aligned}&-\iint_{Q_\tau} [\alpha(z,t)\partial_t\varphi + \beta(z,t)\partial_{zz}\varphi + \gamma(z,t)\partial_z\varphi]\Delta h(z,t)\,dz\,dt\\ &\qquad + \int_0^L [a_1(h_1) - a_1(h_2)]\varphi\Big|_{t=0}^{t=\tau} dz\\ &\qquad + \int_0^\tau [(b_1(h_1) - b_1(h_2))\partial_z\varphi - (\partial_z(b_1(h_1) - b_1(h_2)) - (K_1(h_1) - K_1(h_2)))\varphi]\Big|_{z=0}^{z=L} dt\\ &\quad = \int_0^\tau \Delta K(h_2)(\partial_z h_2 - 1)\varphi\Big|_{z=0}^{z=L} dt\\ &\qquad - \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2 - 1)\partial_z\varphi + \Delta C(h_2)\partial_t h_2\varphi]\,dz\,dt,\end{aligned}$$

where $\alpha(z,t)$, $\beta(z,t)$, and $\gamma(z,t)$ are defined in (1.15), (1.16), and (1.17); i.e., recall that for any smooth function of one variable, F,

$$F(u) - F(v) = G(x)(u-v), \quad \text{where } G(x) = \int_0^1 F'(v(x) + s(u(x) - v(x)))\, ds.$$

If $\varphi(x,t)$ solves the auxiliary problem (1.14), then this expression reduces to (1.18), since in this case

$$\begin{aligned}
[a_1(h_1) - a_1(h_2)]\varphi\Big|_{t=0}^{t=\tau} &= (a_1(h_1) - a_1(h_2))\varphi(z,\tau) \\
&\quad - (a_1(z-1) - a_1(z-1))\varphi(z,0) = 0, \\
[(b_1(h_1) - b_1(h_2))\partial_z\varphi\Big|_{z=0}^{z=L} &= (b_1(0) - b_1(0))\partial_z\varphi(1,t) - (b_1(h_1) - b_1(h_2))\partial_z\varphi(0,t) \\
&= 0 - \Delta h(0,t)\beta(0,t)\partial_z\varphi_2(0,t) = -(p_1(t) - p_2(t))\vartheta_1(t), \\
[(\partial_z(b_1(h_1) - b_1(h_2)) - (K_1(h_1) &- K_1(h_2))) + \Delta K(h_2)(\partial_z h_2 - 1)]\varphi\Big|_{z=0}^{z=L} \\
&= (K_1(h_1)(\partial_z h_1 - 1) - K_2(h_2)(\partial_z h_2 - 1))\varphi\Big|_{z=0}^{z=L} \\
&= \Delta q(t)\vartheta_0(t) - 0.
\end{aligned}$$

The integral identity (1.18) provides an explicit expression relating changes in input to changes in output for the coefficient to data mapping $(p,q) = \Gamma \cdot \Psi_1[C,K]$. Note that for $\varphi = \varphi_0 = \Psi^*[\vartheta_0, 0]$, (1.18) becomes

$$(1.19) \quad \int_0^\tau \Delta p(t)\vartheta_0(t)\, dt = \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z\varphi_0 + \Delta C(h_2)\varphi_0\partial_t h_2]\, dz\, dt,$$

and for $\varphi = \varphi_1 = \Psi^*[0, \vartheta_1]$, (1.18) reduces to the identity

$$(1.20) \quad \int_0^\tau \Delta q(t)\vartheta_1(t)\, dt = \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z\varphi_1 + \Delta C(h_2)\varphi_1\partial_t h_2]\, dz\, dt.$$

The identity (1.18) relates changes ΔC and ΔK in the hydraulic properties of the porous medium to the corresponding changes Δp and Δq which then occur in the measured experimental output for the phase-one experiment. This identity contains information on the Jacobian of the coefficient to data mapping which associates a coefficient pair (C,K) to an output pair (p,q) in the phase-one experiment.

2. The phase-one inverse problem. For admissible coefficients C, K it was shown in Lemma 1.4 that for each t, $0 < t < T$, the solution $h = \Psi_1[C,K]$ of the phase-one direct problem assumes all values between 0, at the bottom of the column, and $p(t) = h(0,t)$ at the top of the column. Then for a fixed $t\varepsilon[0,T]$, the coefficients $C(h(z,t))$ and $K(h(z,t))$ are evaluated at all points of the interval $[p(t), 0]$ as z varies from 0 to L. As (z,t) ranges over the rectangle Q_T, $C(h(z,t))$ and $K(h(z,t))$ are evaluated at all points on the inverval $[-L, 0]$. Then the phase-one experiment explores the interval $[-L, 0]$ in the domain of the coefficients C and K, and the coefficient to data mapping $\Gamma \cdot \Psi$ carries function pairs that are defined and smooth on $[-L, 0]$ onto pairs of data functions (p,q) that are defined and continuous on $[0,T]$. The phase-one inverse problem consists of using the output pair $[p(t), q(t)\colon\ 0 \le t \le T]$ to find the coefficient pair $[C(h), K(h) : -L \le h \le 0]$. Recall that there is no finite time T at which $h(0,T)$ actually equals $-L$ but for every positive ε there exists a finite time

$T = T(\varepsilon)$ such that $h(0,T) = -L+\varepsilon$. Then the data $[p,q]$ on $[0,T]$ determine C and K on $[-L+\varepsilon, 0]$. Since the ε can be chosen to be arbitrarily small, for practical purposes of identification ε is zero; hence the ε will be omitted from subsequent discussions.

The integral identities derived in the previous section suggest a natural interpretation of the sense in which the measured output $[p,q]$ is able to distinguish between coefficient pairs $[C,K]$. It will be shown that the measured outputs p and q determine the coefficients C and K uniquely within an appropriate class of distinguishable functions. Moreover, a constructive algorithm will be defined for inverting the coefficient to data mapping in one such class. Finally, it will be shown that the inverse problem has an equivalent formulation as an output least squares optimization problem whose solution can be shown to solve the inverse problem.

Continuous functions f_1 and f_2 that are not identical on an interval $[\lambda,\rho]$ are said to be *distinguishable* on $[\lambda,\rho]$ if there exists a partition $\{\lambda = \xi_0 < \xi_1 < \cdots < \xi_n = \rho\}$ of $[\lambda,\rho]$ such that on each subinterval (ξ_{m-1},ξ_m) of the partition either

$$\text{(i)} \qquad f_1(x) = f_2(x) \quad \text{for } \xi_{m-1} \le x \le \xi_m$$

or else

$$\text{(ii)} \qquad f_2(x) \ne f_1(x) \quad \text{for } \xi_{m-1} < x < \xi_m.$$

It will be convenient for the proofs to follow if the number of subintervals in the partition is minimal. That is, subintervals on which f_1 coincides with f_2 do not occur consecutively, and consecutive subintervals (ξ_{m-1},ξ_m), (ξ_m,ξ_{m+1}), where the graphs of f_1 and f_2 do not cross, are separated by a point where f_1 equals f_2; i.e., $f_1(\xi_m) = f_2(\xi_m)$. A partition with this property will be said to be *adjusted to* the functions f_1 and f_2.

Functions that are not distinguishable need not be equal at all points of their domain. For example, the function $f_1(x) = x\sin(1/x)$ is neither distinguishable from nor identical to $f_2(x) = 0$ on $[0,1]$, since f_1 is clearly not identical to f_2 yet f_1 and f_2 are equal at infinitely many points in every neighborhood of zero. However, there exist classes of functions which are proper subsets of the continuous functions and any two functions from the class are either distinguishable on the interval of definition or else they are identical there. Analytic functions are one such class, as are the so-called polygonal functions that are continuous and piecewise linear on their interval of definition.

THEOREM 2.1. *For admissible coefficients C_j and K_j, $j = 1,2$, let $h_j(z,t) = \Psi_1[C_j,K_j]$. Let $(p_j,q_j) = \Gamma \cdot \Psi_1[C_j,K_j]$ for $j = 1,2$. If the pairs C_1,K_1 and C_2,K_2 are distinguishable on the interval $[-L,0]$ then p_1,p_2 and q_1,q_2 are not identical on $[0,T]$.*

Proof. Suppose that C_1,K_1 and C_2,K_2 are distinguishable on the interval $[-L,0]$ and $q_1 = q_2$ and $p_1 = p_2$ on $[0,T]$. It will now be shown that the two conditions are inconsistent.

Let $\{0 = \xi_0 > \xi_1 > \cdots > \xi_n = h(0,T)\}$ denote a partition of $[h(0,T),0]$ adjusted to the distinguishable functions K_1 and K_2 and let $\{0 = \xi_0 > \xi_1' > \cdots \xi_{n-1}' > \xi_n\}$ denote a (possibly different) partition of $[h(0,T),0]$ adjusted to C_1 and C_2. For the sake of discussion, suppose that $K_1(h) \ne K_2(h)$ for $\xi_0 < h < \xi_1$ and suppose $C_1(h) \ne C_2(h)$ for $\xi_0 < h < \xi_1'$. Let η_1 denote the smaller of the two numbers ξ_1 and ξ_1' and let $\tau \le T$ denote the unique value at which the monotone function $q(t)$ satisfies $q(\tau) = \eta_1$. Then $\Delta K(h) = K_1(h) - K_2(h)$ and $\Delta C(h) = C_1(h) - C_2(h)$ do not vanish at any point of the interval $(0,\eta_1)$.

Let $\varphi_1 = \Psi^*[0,\vartheta_1]$, where the input function $\vartheta_1(t)$ satisfying $\vartheta_1(\tau) = 0$ is chosen to be sufficiently large and positive for $0 < t < \tau$ that

$$\vartheta_1(t) > \varphi_1(z,t) > 0 \quad \text{and} \quad \partial_z \varphi_1(z,t) > 0 \quad \text{for } 0 < z < L, \quad 0 < t < \tau. \tag{2.1}$$

Similarly, for $\vartheta_0(t)$ satisfying $\vartheta_0(\tau) = 0$, choose ϑ_0 sufficiently large and positive for $0 < t < \tau$ that $\varphi_2 = \Psi^*[\vartheta_0, 0]$ satisfies

$$\varphi_2(z,t) < 0 \quad \text{and} \quad \partial_z \varphi_2(z,t) > 0 \quad \text{for } 0 < z < L, \quad 0 < t < \tau. \tag{2.2}$$

It follows from Lemmas 1.2 and 1.3 that $\partial_z h_2(z,t) - 1$ and $\partial_t h_2(z,t)$ are strictly negative almost everywhere on Q_τ. If $q_1 = q_2$ and $p_1 = p_2$ on $[0,T]$ then $\Delta q(t) = q_1(t) - q_2(t) = 0$ and $\Delta p(t) = p_1(t) - p_2(t) = 0$ for $0 \leq t \leq \tau$, and it follows from (1.19) and (1.20) that

$$\iint_{Q_\tau} \Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z \varphi_1 \, dz \, dt = -\iint_{Q_\tau} \Delta C(h_2)\varphi_1 \partial_t h_2 \, dz \, dt, \tag{2.3}$$

$$\iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z \varphi_2 \, dz \, dt = -\iint_{Q_\tau} \Delta C(h_2)\varphi_2 \partial_t h_2 \, dz \, dt. \tag{2.4}$$

Now because of (2.1), (2.3) implies that ΔK and ΔC have opposite signs, whereas (2.4), taken with (2.2), implies ΔK and ΔC are of the same sign on $(0, \eta_1)$. It then follows that q_1, q_2 and p_1, p_2 cannot be identical on $[0,T]$ if C_1, C_2 and K_1, K_2 are distinguishable on the interval $[-L, 0]$. Only a slight modification of the argument is required to deal with alternative possibilities for ΔC and ΔK on the initial subintervals of the partitions.

Inverse problems are often reformulated as optimization problems in which an error functional based on the measured output data is to be minimized over a class of admissible inputs. Although this has the advantage of providing a means of computing a solution to the problem, it is often not evident that the solution of the optimization problem is also a solution of the inverse problem. It will be shown here that for a properly formulated error functional, any pair of admissible coefficients (C,K) that minimizes the error functional must also solve the inverse problem.

Theorem 2.2. *For a fixed pair of admissible coefficients C_0, K_0, let $(p(t; C_0, K_0)$, $q(t; C_0, K_0)) = \Gamma \cdot \Psi_1[C_0, K_0]$ denote the corresponding measured output. For an arbitrary pair of admissible coefficients C, K, let $h(z,t) = \Psi_1[C,K]$ and $(p(t; C, K)$, $q(t; C, K)) = \Gamma \cdot \Psi_1[C,K]$. Define the output least squares error functional associated to the pair (C,K) to be*

$$J[(C,K)] = \int_0^T (p(t; C,K) - p(t; C_0, K_0)^2 dt + \int_0^T (q(t; C,K) - q(t; C_0, K_0))^2 dt. \tag{2.5}$$

Then the variation of this functional is given by

$$\delta J[(C,K), (\delta C, \delta K)] = \iint_{Q_T} [\delta K(h)(\partial_z h(z,t) - 1)\partial_z \varphi + \delta C(h)\partial_t h(z,t)\varphi] \, dz \, dt, \tag{2.6}$$

where $\varphi(z,t) = \Psi^*[\vartheta_0, \vartheta_1]$ with data $\vartheta_0(t) = 2(p(t; C,K) - p(t; C_0, K_0))$ and $\vartheta_1(t) = 2(q(t; C,K) - q(t; C_0, K_0))$. Moreover, if (C,K) minimizes the functional J then $\Gamma \cdot \Psi_1[C,K] = \Gamma \cdot \Psi_1[C_0, K_0]$; i.e., (C,K) solves the inverse problem.

Proof. The variation of the functional J is easily computed to be

$$\delta J[(C,K), (\delta C, \delta K)] = \int_0^T 2(p(t; C,K) - p(t; C_0, K_0))\delta p(t) \, dt + \int_0^T 2(q(t; C,K) - q(t; C_0, K_0))\delta q(t) \, dt,$$

where δp and δq denote the differences $\delta p(t) = p(t; C+\delta C, K+\delta K) - p(t; C, K)$ and $\delta q(t) = q(t; C+\delta C, K+\delta K) - q(t; C, K)$. If φ denotes the solution of (1.14) for data $\vartheta_0(t) = 2(p(t; C, K) - p(t; C_0, K_0))$ and $\vartheta_1(t) = 2(q(t; C, K) - q(t; C_0, K_0))$, then the result (1.18) with $h_2(z,t) = \Psi_1[C, K]$ asserts that

$$\begin{aligned}\delta J[(C,K),(\delta C,\delta K)] &= \int_0^T [\vartheta_0(t)\delta p(t) + \vartheta_1(t)\delta q(t)]\,dt \\ &= \iint_{Q_T} [\delta C(h_2)\varphi\partial_t h_2 + \delta K(h_2)(\partial_z h_2 - 1)\partial_z\varphi]\,dz\,dt.\end{aligned}$$

But this is precisely (2.6).

Now if (C, K) is an admissible coefficient pair that causes the variation $\delta J[(C,K),(\delta C,\delta K)]$ to vanish for arbitrary perturbations $(\delta C, \delta K)$ of the coefficients, then it follows from (2.6) in combination with Lemmas 1.2 and 1.3 that φ and $\delta_z\varphi$ must vanish on Q_T. However, φ is the solution of the parabolic problem (1.14) and can vanish identically if and only if the data $\vartheta_0(t)$ and $\vartheta_1(t)$ are each zero. But this is to say that the computed outputs $p(t; C, K)$ and $q(t; C, K)$ agree with the measured outputs $p(t; C_0, K_0)$ and $q(t; C_0, K_0)$. Then (C, K) solves the inverse problem.

This result establishes that if the inverse problem is reformulated as an output least squares problem based on the overspecified data p and q, then a solution of the output least squares problem is truly a solution of the inverse problem.

Theorem 2.1 asserts that distinguishable coefficient pairs cannot produce pairs of data functions that are identical; i.e., the measured boundary data (p, q) determine the coefficient pair (C, K) uniquely within any subset of the admissible coefficients where functions that are indistinguishable are identical. One such subset is the set of polygonal coefficient functions constructed on a fixed partition of $[-L, 0]$.

Theorem 2.1 may be interpreted as an assertion that the coefficient to data mapping $(p, q) = \Gamma\cdot\Psi_1[C, K]$ is injective from the set of polygonal coefficients to admissible data pairs, but this is not quite sufficient to ensure invertibility. In fact, the coefficient to data mapping is separately monotone in each coefficient and a monotone mapping on a totally ordered space is invertible. Hence if the space of coefficients can be parameterized in a totally ordered fashion then the coefficient to data mapping can be inverted. Such a parameterization and inversion becomes possible due to the fact that monotonicity permits the coefficient to data mapping to be factored into a product of simpler maps.

Let p and q denote admissible data functions on an interval $[0, T]$, and let $\{0 = t_0 < \cdots < t_n = T\}$ denote a partition of $[0, T]$. The data function $p(t) = h(0, t)$ is monotone decreasing from the value 0 at $t = 0$ down to the value $p(T) = -L$ (i.e., $-L + \varepsilon$) at $t = T$. Then the data values $p_m = p(t_m)$ define a corresponding partition, $\{0 = p_0 > \cdots > p_n = -L\}$ of the interval $[-L, 0]$, which is the domain of the coefficients C, K during phase one.

Define a mapping Π_n: $\mathbb{C}[-L, 0] \longrightarrow \mathbb{R}^{n+1}$ which carries functions $K(h)$ and $C(h)$ from $\mathbb{C}[-L, 0]$ onto arrays of values $\{\kappa_m\}$ and $\{\xi_n\}$ equal to the values of the coefficients at the node points of the partition $\{p_m\colon\ m = 0, \ldots, n\}$ of their domain $[-L, 0]$; i.e.,

(2.7)
$$\Pi_n K = \{\kappa_m = K(p_m) : 0 \le m \le n\} \quad \text{and} \quad \Pi_n C = \{\xi_m = C(p_m) : 0 \le m \le n\}.$$

It will also be convenient to define a mapping P_n carrying an array of $n+1$ real values to a polygonal (continuous and piecewise-linear) function. This mapping, P_n, may be defined to have the following action on the array $\{\kappa_0, \kappa_1, \ldots, \kappa_n\}$:

$$P_n\{\kappa_m\} = \sum_{m=0}^{n} \kappa_m \Lambda_m(v), \tag{2.8}$$

where $\Lambda_m(v)$ denotes the piecewise-linear, continuous "hat functions" defined as follows on the partition $\{p_m\}$ of $[-L, 0]$: for $m = 0, 1, 2, \ldots, n$,

$$\Lambda_0(v) = 1 - v/p_1 \qquad \text{if } 0 = p_0 \le v \le p_1,$$

$$\Lambda_m(v) = \begin{cases} \dfrac{v - p_{m-1}}{p_m - p_{m-1}}, & \text{if } p_{m-1} \le v \le p_m, \\ \dfrac{p_{m+1} - v}{p_{m+1} - p_m} & \text{if } p_m \le v \le p_{m+1}, \end{cases}$$

$$\Lambda_m(v) = 0 \qquad \text{if } v < p_{m-1} \quad \text{or} \quad v > p_{m+1}.$$

For admissible coefficients C and K and a given partition $\{p_m\}$ of $[-L, 0]$, the polygonal functions $C^* = P_n\Pi_n C$ and $K^* = P_n\Pi_n K$ denote the unique polygonal approximations to C and K on the partition $\{p_m\}$; i.e., at each node point $p = p_m$, $C^*(p) = C(p)$ and $K^*(p) = K(p)$.

Similarly, for arbitrary arrays $\{\kappa_0, \kappa_1, \ldots, \kappa_n\}$ and $\{\xi_0, \ldots, \xi_n\}$ of positive parameters, (2.8) defines polygonal coefficients $K^*(v) = P_n[\kappa_0, \ldots, \kappa_n](v)$ and $C^*(v) = P_n[\xi_0, \ldots, \xi_n](v)$ on $p_n \le v \le p_0 = 0$. When these coefficients are used, the direct problem (1.1) has a unique solution $h^* = \Psi_1[C^*, K^*]$ on Q_T[9]. The corresponding data pair $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ is an admissible pair by the lemmas of the previous section. An algorithm can now be defined that produces arrays $\{\kappa_0, \kappa_1, \ldots, \kappa_n\}$ and $\{\xi_0, \ldots, \xi_n\}$ of parameters for which the corresponding computed output (p^*, q^*) approximates the measured data (p, q).

Assume an admissible data pair (p, q) is given on $[0, T]$. Let $\{t_m : 0 \le m \le n\}$ denote an arbitrary partition of $[0, T]$, and let $\{p_m : 0 \le m \le n\}$ denote the associated partition of $[p(T), 0]$ induced by the monotone function $p(t)$. For each k, $1 \le k \le n$, let $\varphi_k(z, t) = \Psi^*[\vartheta_k, 0]$ and $\psi_k(z, t) = \Psi^*[0, \rho_k]$ denote solutions to adjoint problems (1.14) with $\tau = t_k$. For each k, (1.14) is equivalent to a linear one-dimensional parabolic initial boundary value problem that is driven from a zero initial state by controlling the flux $\vartheta_k(t)$ at the left endpoint or by controlling the function value $\rho_k(t)$ at the right endpoint. The data functions ϑ_k and ρ_k in the adjoint problems are chosen to satisfy $\vartheta_k(t_k) = \rho_k(t_k) = 0$. In addition, $\vartheta_k(t)$, $\rho_k(t)$ can be chosen sufficiently large and positive for $0 < t < t_k$ that it follows

$$\varphi_k(z, t) < 0, \qquad \partial_z \varphi_k(z, t) > 0,$$

and

$$\psi_k(z, t) > 0, \qquad \partial_z \psi_k(z, t) > 0 \quad \text{on } Q_\tau.$$

Then the pairs $(\kappa_0, \xi_0), (\kappa_1, \xi_1), \ldots, (\kappa_n, \xi_n)$ of parameters defining the polygonal approximations for K and C are constructed, one at a time, according to the following algorithm.

Step 0. Assume $\kappa_0 = K(0)$ and $\xi_0 = C(0)$ are given. (These are the values of K and C at saturation.)

Step 1. For $\tau = t_1$, let

$$\begin{aligned} C^\sharp &= P_n[\xi_0, \xi_0, \dots, \xi_0] \quad \text{and} \quad K^\sharp = P_n[\kappa_0, \kappa_0, \dots, \kappa_0], \\ h_2(z,t) &= \Psi_1[C^\sharp, K^\sharp] \\ (p^\sharp, q^\sharp) &= \Gamma \cdot \Psi_1[C^\sharp, K^\sharp]. \end{aligned}$$

Then (κ_1, ξ_1) are given by

$$\begin{pmatrix} \kappa_1 \\ \xi_1 \end{pmatrix} = \begin{pmatrix} \kappa_0 \\ \xi_0 \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_1 \\ e_1 \end{pmatrix}, \tag{2.9}$$

where the entries of the 2×2 matrix M are given by

$$\begin{aligned} M_{11} &= \iint_{Q_\tau} \Lambda_1(h_2)(\partial_z h_2 - 1)\partial_z \varphi_1 \, dz \, dt, \qquad & M_{12} &= \iint_{Q_\tau} \Lambda_1(h_2)\varphi_1 \partial_t h_2 \, dz \, dt, \\ M_{21} &= \iint_{Q_\tau} \Lambda_1(h_2)(\partial_z h_2 - 1)\partial_z \psi_1 \, dz \, dt, \qquad & M_{22} &= \iint_{Q_\tau} \Lambda_1(h_2)\psi_1 \partial_t h_2 \, dz \, dt \end{aligned} \tag{2.10}$$

and

$$d_1 = \int_0^T (q - q^\sharp)\vartheta_1(t) \, dt, \qquad e_1 = \int_0^\tau (p - p^\sharp)\rho_1(t) \, dt. \tag{2.11}$$

Then $C^*(h) = \xi_0 \Lambda_0(h) + \xi_1 \Lambda_1(h)$ and $K^*(h) = \kappa_0 \Lambda_0(h) + \kappa_1 \Lambda_1(h)$ for $p_1 \leq h \leq 0$.

Proceed in this way to generate pairs $(\kappa_1, \xi_1), \dots, (\kappa_{m-1}, \xi_{m-1})$ which define the polygonal coefficients C^* and K^* on the portion $p_{m-1} \leq h \leq 0$ of their domain. Pairs $(\kappa_i, \xi_i), i = 0, 1, \dots, m-1$, are known at this stage. To compute the next pair (ξ_m, κ_m), given the pairs (κ_i, ξ_i), $0 \leq i \leq m-1$, continue as follows:

Step m. For $\tau = t_m$, let

$$\begin{aligned} C^\sharp &= P_n[\xi_0, \xi_1, \dots, \xi_{m-1}, \dots, \xi_{m-1}] \text{ and } K^\sharp = P_n[\kappa_0, \kappa_1, \dots, \kappa_{m-1}, \dots, \kappa_{m-1}], \\ h_2(z,t) &= \Psi_1[C^\sharp, K^\sharp] \\ (p^\sharp, q^\sharp &= \Gamma \cdot \Psi_1[C^\sharp, K^\sharp]. \end{aligned}$$

Then (κ_m, ξ_m) are given by

$$\begin{pmatrix} \kappa_m \\ \xi_m \end{pmatrix} = \begin{pmatrix} \kappa_{m-1} \\ \xi_{m-1} \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_{m-1} \\ e_{m-1} \end{pmatrix}, \tag{2.12}$$

where the entries of the 2×2 matrix M are given by

$$\begin{aligned} M_{11} &= \iint_{Q_\tau} \Lambda_1(h_2)(\partial_z h_2 - 1)\partial_z \varphi_m \, dz \, dt, \qquad & M_{12} &= \iint_{Q_\tau} \Lambda_1(h_2)\varphi_m \partial_t h_2 \, dz \, dt, \\ M_{21} &= \iint_{Q_\tau} \Lambda_1(h_2)(\partial_z h_2 - 1)\partial_z \psi_m \, dz \, dt, \qquad & M_{22} &= \iint_{Q_\tau} \Lambda_1(h_2)\psi_m \partial_t h_2 \, dz \, dt \end{aligned} \tag{2.13}$$

and

$$d_1 = \int_0^\tau (q - q^\sharp)\vartheta_m(t) \, dt, \qquad e_1 = \int_0^\tau (p - p^\sharp)\rho_m(t) \, dt. \tag{2.14}$$

In this way, n pairs of parameter values can be generated. It remains to be seen in what sense the polygonal coefficients based on these parameter pairs can produce computed output that approximates the measured output (p, q) used in the algorithm to find the parameters.

THEOREM 2.3. *Let* $[p(t), q(t)]$ *denote an admissible data pair on* $[0, T]$*, and let* $\{0 = t_0 < \cdots < t_n = T\}$ *denote an arbitrary partition of the interval* $[0, T]$*. Suppose that parameter pairs* $\{(\xi_0, \kappa_0), \ldots, (\xi_n, \kappa_n)\}$ *have been generated using the algorithm described previously and that* $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ *for* $C^* = P_n[\xi_0, \ldots, \xi_n]$ *and* $K^* = P_n[\kappa_0, \ldots, \kappa_n]$*. Then*

(a) $\int_0^{t_m}(p(t) - p^*(t))\vartheta_m(t)\,dt = \int_0^{t_m}(q(t) - q^*(t))\rho_m(t)\,dt = 0, m = 1, \ldots, n,$

(b) *if* $p, q \varepsilon \mathbb{C}^2[t_{m-1}, t_m]$ *for* $m = 1, \ldots n$*, then* $|p(t_m) - p^*(t_m)| \leq \nu\Delta t^2$ *and* $|q(t_m) - q^*(t_m)| \leq \nu\Delta t^2$ *for* $m = 1, \ldots, n$ *for a positive constant* ν *depending on* p *and* q *and* $\Delta t = \max(t_m - t_{m-1})$.

Proof. Suppose (p, q) is an admissible data pair; i.e., p and q satisfy (1.13) and (1.4), respectively. Let $\{0 = t_0 < \cdots < t_n = T\}$ denote a partition of the interval $[0, T]$, and let $\{0 = p_0 > \cdots > p_n = -L\}$ denote the corresponding partition of the interval $[-L, 0]$ induced by the monotone function, $p(t)$.

The initial parameter pair $\xi_0 = C(0)$, $\kappa_0 = K(0)$ is assumed to be known. This represents no loss of generality since these values are often known from independent experiments, or they may be determined by asymptotic estimates from the data $p(t)$, $q(t)$, [5, 6].

The computed outputs $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ have been generated by the polygonal coefficients $C^* = P_n[\xi_0, \ldots, \xi_n]$ and $K^* = P_n[\kappa_0, \ldots, \kappa_n]$. Let $(p^\sharp, q^\sharp)$ denote outputs computed from the coefficients $C^\sharp = P_n[\xi_0, \ldots, \xi_0]$ and $K^\sharp = P_n[\kappa_0, \ldots, \kappa_0]$, and let

$$\begin{pmatrix} \kappa_1' \\ \xi_1' \end{pmatrix} = \begin{pmatrix} \kappa_0 \\ \xi_0 \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_1' \\ e_1' \end{pmatrix}, \tag{2.15}$$

where the entries of the 2×2 matrix M are given by (2.10) and

$$d_1' = \int_0^{t_1} (q - q^*)\vartheta_1(t)\,dt, \qquad e_1' = \int_0^{t_1} (p - p^*)\rho_1(t)\,dt. \tag{2.16}$$

Then it follows from (2.9), (2.11) and (2.15), (2.16) that

$$\begin{pmatrix} \kappa_1' \\ \xi_1' \end{pmatrix} = \begin{pmatrix} \kappa_0 \\ \xi_0 \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_1 \\ e_1 \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_1'' \\ e_1'' \end{pmatrix},$$

where

$$d_1'' = \int_0^{t_1} (q^\sharp - q^*)\vartheta_1(t)\,dt, \qquad e_1'' = \int_0^{t_1} (p^\sharp - p^*)\rho_1(t)\,dt.$$

This may be written as

$$\begin{pmatrix} \kappa_1' \\ \xi_1' \end{pmatrix} = \begin{pmatrix} \kappa_1 \\ \xi_1 \end{pmatrix} + [M]^{-1} \begin{pmatrix} d_1'' \\ e_1'' \end{pmatrix}. \tag{2.17}$$

Applying the integral identities (1.19) and (1.20) to the coefficient pairs $(C_1, K_1) = (C^\sharp, K^\sharp)$ and $(C_2, K_2) = (C^*, K^*)$ leads to

$$\begin{aligned} (\kappa_0 - \kappa_1)M_{11} + (\xi_0 - \xi_1)M_{12} &= d_1'', \\ (\kappa_0 - \kappa_1)M_{21} + (\xi_0 - \xi_1)M_{22} &= e_1''. \end{aligned}$$

That is,

$$[M]\begin{pmatrix}\kappa_0-\kappa_1\\ \xi_0-\xi_1\end{pmatrix}=\begin{pmatrix}d_1''\\ e_1''\end{pmatrix}, \tag{2.18}$$

where the entries of the 2×2 matrix M are given by (2.10) and we have used the fact that $\Delta K(v)=K^\sharp(v)-K^*(v)=\kappa_0\Lambda_0(v)+\kappa_0\Lambda_1(v)-(\kappa_0\Lambda_0(v)+\kappa_1\Lambda_1(v))=(\kappa_0-\kappa_1)\Lambda_1(v)$ and $\Delta C(v)=(\xi_0-\xi_1)\Lambda_1(v)$. Now (2.17) and (2.18) together imply that

$$\begin{pmatrix}\kappa_1'\\ \xi_1'\end{pmatrix}=\begin{pmatrix}\kappa_1\\ \xi_1\end{pmatrix}+[M]^{-1}[M]\begin{pmatrix}\kappa_0-\kappa_1\\ \xi_0-\xi_1\end{pmatrix}=\begin{pmatrix}\kappa_0\\ \xi_0\end{pmatrix},$$

and this result, taken with (2.15), implies $d_1'=e_1'=0$. However, this is result (a) in the case where $m=1$. In much the same way, one can show that (a) holds for $1<m\le n$ by letting

$$\begin{pmatrix}\kappa_m'\\ \xi_m'\end{pmatrix}=\begin{pmatrix}\kappa_{m-1}\\ \xi_{m-1}\end{pmatrix}+[M]^{-1}\begin{pmatrix}d_{m-1}'\\ e_{m-1}'\end{pmatrix}$$

and proceeding as above to show that $d_{m-1}'=e_{m-1}'=0$.

To prove (b) introduce the family of functions $\{\lambda_m(t)\}$, piecewise-linear, continuous functions on the partition $\{t_m\}$ of $[0,T]$. The family $\{\lambda_m(t)\}$ is analogous to the family $\{\Lambda_m(v)\}$ on the partition $\{p_m\}$ of $[-L,0]$. Then, with a slight abuse of notation, $P_nq(t)=[q(t_0),q(t_1),\ldots,q(t_n)]\varepsilon\mathbb{R}^{n+1}$ and

$$\Pi_nP_nq(t)=\sum_{i=0}^{n}q(t_i)\lambda_i(t),$$

where Π_nP_nq denotes the unique polygonal approximation to $q(t)$ on the partition $\{t_m\}$ of $[0,T]$.

For outputs p^* and q^* generated by the polygonal coefficients of the algorithm, use (a) in the case where $m=1$ to write

$$\int_0^{t_1}p(t)-p^*(t))\vartheta_1(t)\,dt=\int_0^{t_1}(p(t)-\Pi_nP_np(t))\vartheta_1(t)\,dt$$
$$+\int_0^{t_1}(\Pi_nP_np(t)-\Pi_nP_np^*(t))\vartheta_1(t)\,dt+\int_0^{t_1}(\Pi_nP_np^*(t)-p^*(t))\vartheta_1(t)\,dt=0.$$

On $[0,t_1]$, $\Pi_nP_np(t)-\Pi_nP_np^*(t)=(p(t_1)-p^*(t_1))\lambda_1(t)$; hence

$$|p(t_1)-p^*(t_1)|=\left|\int_0^{t_1}(p(t)-\Pi_nP_np(t))\vartheta_1(t)\,dt+\int_0^{t_1}(\Pi_nP_np^*(t)-p^*(t))\vartheta_1(t)\,dt|\right|/I_1,$$

where $I_1=\int_0^{t_1}\lambda_1(t)\vartheta_1(t)\,dt>0$. Making use of a well-known property of the trapezoidal approximation for integrals, it follows that if $p\varepsilon\mathbb{C}^2[0,t_1]$ then for η_1, a positive constant depending on $p(t)$ and $p^*(t)$,

$$|p(t_1)-p^*(t_1)|\le\eta_1(t_1-0)^3=\eta_1\Delta t_1^3.$$

Similarly, $|q(t_1)-q^*(t_1)|\le\eta_1\Delta t_1^3$ with a possibly large value of η_1.

For $m = 2$, it follows from (a) that

$$\int_{t_1}^{t_2} (\Pi_n P_n p^*(t) - \Pi_n P_n p(t))\vartheta_2(t)\,dt = \int_0^{t_2} (p(t) - \Pi_n P_n p(t))\vartheta_2(t)\,dt$$
$$+ \int_0^{t_2} (\Pi_n P_n p^*(t) - p^*(t))\vartheta_2(t)\,dt - \int_0^{t_1} (\Pi_n P_n p^*(t) - \Pi_n P_n p(t))\vartheta_2(t)\,dt.$$

Hence

$$|p(t_2) - p^*(t_2) \leq \left|\int_0^{t_2} (p(t) - \Pi_n P_n p(t))\vartheta_2(t)\,dt\right| / I_2$$
$$+ \left|\int_0^{t_2} (\Pi_n P_n p^*(t) - p^*(t))\vartheta_2(t)\,dt\right| / I_2 + \eta_1 \Delta t_1^3 \int_0^{t_2} \lambda_1(t)\vartheta_2(t)\,dt / I_2.$$

Again using the property of the trapezoidal approximation for integrals leads to

$$|p(t_2) - p^*(t_2)| \leq \eta_1 \Delta t_1^3 + \eta_2 \Delta t_2^3 + \eta_1^* \Delta t_1^3 + \eta_2^* \Delta t_2^3 + \eta_1 I_1' / I_2 \Delta t_1^3.$$

Then there exist positive constants, which we denote again by η_1 and η_2, such that

$$|p(t_2) - p^*(t_2)| \leq \eta_1 \Delta t_1^3 + \eta_2 \Delta t_2^3 \quad \text{and} \quad |q(t_2) - q^*(t_2)| \leq \eta_1 \Delta t_1^3 + \eta_2 \Delta t_2^3.$$

For each m, $1 \leq m \leq n$, one can show in this way that

$$|p(t_m) - p^*(t_m)| \leq \sum_{i=1}^{m} \eta_i \Delta t_i^3 \quad \text{and} \quad |q(t_m) - q^*(t_m)| \leq \sum_{i=1}^{m} \eta_i \Delta t_i^3.$$

Letting $\Delta t = \max\{\Delta t_i\colon 1 \leq i \leq n\}$, note that $m\Delta t \leq n\Delta t \sim T$; hence if $\eta = \max \eta_i$, then

$$|p(t_m) - p^*(t_m)| \leq (\eta n \Delta t)\Delta t^2 \quad \text{and} \quad |q(t_m) - q^*(t_m)| \leq (\eta n \Delta t)\Delta t^2.$$

Letting $\nu = \eta T$ gives part (b) of the theorem.

Theorem 2.3 shows that given any admissible data pair (p, q) it is always possible to construct a pair of polygonal coefficients C^*, K^* such that the resulting computed output $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ reproduces the measured output (p, q) with accuracy dependent on the mesh size for the partition $\{t_m\}$. In case it is known that the measured data (p, q) are in fact equal to $\Gamma \cdot \Psi_1[C, K]$ for some admissible coefficient pair (C, K), then the following theorem describes how (C^*, K^*) is related to (C, K).

THEOREM 2.4. *Let $(p, q) = \Gamma \cdot \Psi_1[C, K]$ for (C, K), an admissible coefficient pair, and let $\{0 = t_0 < \cdots < t_n = T\}$ denote an arbitrary partition of the interval $[0, T]$. Suppose parameter pairs $\{(\xi_0 \kappa_0), \ldots, (\xi_n, \kappa_n)\}$ have been generated using the algorithm described previously and that $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ for $C^* = P_n[\xi_0, \ldots, \xi_n]$ and $K^* = P_n[\kappa_0, \ldots, \kappa_n]$. If $C, K \varepsilon \mathbb{C}^2[p_i, p_{i-1}]$ for $1 \leq i \leq n$, then*

$$|\xi_i - C(p_i)| \leq \nu \Delta p \quad \text{and} \quad |\kappa_i - K(p_i)| \leq \nu \Delta p \quad \text{for } 1 \leq i \leq n,$$

for a positive constant ν depending on C and K for $\Delta p = \max\{p_{i-1} - p_i\}$.

Proof. Apply the integral identities (1.19) and (1.20) with the coefficient pairs $(C_1, K_1) = (C, K)$ and $(C_2, K_2) = (C^*, K^*)$ and $\tau = t_1$. Then with $\varphi_j = \Psi^*[\vartheta_j, 0]$

and $\psi_j = \Psi^*[0,\rho_j]$ as they were in Theorem 2.3, by part (a) of Theorem 2.3,

$$\iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z \varphi + \Delta C(h_2)\varphi\partial_t h_2]\, dz\, dt = 0,$$
$$\iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z,t) - 1)\partial_z \psi + \Delta C(h_2)\psi\partial_t h_2]\, dz\, dt = 0,$$

where $\Delta K(h) = K(h) - K^*(h)$ and $\Delta C(h) = C(h) - C^*(h)$. Now letting $\Pi_n P_n K$ and $\Pi_n P_n C$ denote the polygonal approximations to K and C on the partition $\{p_m\}$ of $[-L, 0]$ leads to

$$\begin{aligned}\iint_{Q_\tau} &(\Pi_n P_n K(h_2) - K^*(h_2))(\partial_z h_2(z,t) - 1)\partial_z \varphi\, dz\, dt \\ &+ \iint_{Q_\tau} (\Pi_n P_n C(h_2) - C^*(h_2))\varphi\partial_t h_2\, dz\, dt \\ &= \iint_{Q_\tau} (\Pi_n P_n K(h_2) - K(h_2))(\partial_z h_2(z,t) - 1)\partial_z \varphi\, dz\, dt \\ &+ \iint_{Q_\tau} (\Pi_n P_n C(h_2) - C(h_2))\varphi\partial_t h_2\, dz\, dt,\end{aligned}$$

$$\begin{aligned}\iint_{Q_\tau} &(\Pi_n P_n K(h_2) - K^*(h_2))(\partial_z h_2(z,t) - 1)\partial_z \psi\, dz\, dt \\ &+ \iint_{Q_\tau} (\Pi_n P_n C(h_2) - C^*(h_2))\psi\partial_t h_2\, dz\, dt \\ &= \iint_{Q_\tau} (\Pi_n P_n K(h_2) - K(h_2))(\partial_z h_2(z,t) - 1)\partial_z \psi\, dz\, dt \\ &+ \iint_{Q_\tau} (\Pi_n P_n C(h_2) - C(h_2))\psi\partial_t h_2\, dz\, dt.\end{aligned}$$

It follows from Lemma 1.4 that $p_1 \le u_2(z,t) \le p_0 = 0$ for $(z,t)\varepsilon Q_{t_1}$. Then

$$\begin{aligned}\Pi_n P_n K(h_2) - K^*(h_2) &= (K(p_1) - \kappa_1)\Lambda_1(u_2(z,t)), \\ \Pi_n P_n C(h_2) - C^*(h_2) &= (C(p_1) - \xi_1)\Lambda_1(u_2(z,t)) \quad \text{on } Q_{t_1}\end{aligned}$$

and for M the 2×2 matrix whose entries are given by (2.10)

$$[M]\begin{bmatrix} K(p_1) - \kappa_1 \\ C(p_1) - \xi_1 \end{bmatrix} = \begin{bmatrix} I(K,\varphi) + J(C,\varphi) \\ I(K,\varphi) + J(C,\psi) \end{bmatrix}. \tag{2.19}$$

Using the lemmas of section 1 together with a well-known property of the polygonal approximation for a smooth function, it follows that for $\tau = t_1$ there exists a positive constant ν_1 such that

$$|I(K,\varphi)| = \left| \iint_{Q_\tau} (\Pi_n P_n K(h_2) - K(h_2))(\partial_z h_2(z,t) - 1)\partial_z \varphi\, dz\, dt \right| \le \nu_1 (p_1 - p_0)^2,$$
$$|J(C,\varphi)| = \left| \iint_{Q_\tau} (\Pi_n P_n C(h_2) - C(h_2))\varphi\partial_t h_2\, dz\, dt \right| \le \nu_1 (p_1 - p_0)^2,$$

and for a larger constant, which we still denote by ν_1,

$$|I(K,\varphi)| + |J(C,\varphi)| \le \nu_1 (p_1 - p_0)^2 \quad \text{and} \quad |I(K,\psi)| + |J(C,\psi)| \le \nu_1 (P_1 - p_0)^2.$$

Then it follows from (2.19) that there exists a positive constant, denoted again by ν_1, such that

$$|K(p_1) - \kappa_1| \le \nu_1 (p_1 - p_0)^2 \quad \text{and} \quad |C(p_1) - \xi_1| \le \nu_1 (p_1 - p_0)^2.$$

Continuing in a fashion similar to what was done in the proof of Theorem 2.3, we obtain for each i, $1 \le i \le n$,

$$|K(p_i) - \kappa_i| \le \sum_{j=1}^{i} \nu_j \Delta p_j^2 \quad \text{and} \quad |C(p_i) - \xi_j| \le \sum_{j=1}^{i} \nu_j \Delta p_j^2.$$

Let Δp and ν denote the largest of the numbers Δp_j and ν_j, respectively, and note that $n\Delta p \sim L$. Then for each i, $1 \le i \le n$,

$$|K(p_i) - \kappa_i| \le L\nu\Delta p \quad \text{and} \quad |C(p_i) - \xi_i| \le L\nu\Delta p,$$

which is what was to be proven.

Theorem 2.4 asserts that for a vertical column, initially saturated and allowed to drain to equilibrium under the force of gravity, it is possible to construct polygonal functions which approximate the hydraulic functions $C(h)$, $K(h)$ on the range $-L \le h \le 0$ for a column of length L; i.e., this phase-one experiment determines the hydraulic coefficients on the range $-L \le h \le 0$ from data $\{p(t), q(t)\colon\ 0 \le t \le T\}$ measured at the top and bottom of the column. The monotonicity of the data function $p(t)$ permits the parameter pairs characterizing the polygonal functions approximating C and K to be found one at a time rather than as an ensemble. In effect, the coefficient to data mapping is factored into n mappings, each of which associates a coefficient parameter pair (ξ_i, κ_i) to a data parameter pair (p_i, q_i). Each of these factor mappings is then uniquely invertible as expressed, for example in (2.12).

It will be shown in the next section that the range over which the coefficients can be determined can be extended by applying suction to the bottom of the column in what will be referred to as phase two of the experiment.

3. The phase-two direct problem. Consider a vertical soil column, initially saturated and allowed to drain to a state of equilibrium under gravity. The column will now be drained further by applying suction. If there is no flow across the top end of the column and a suction $s(t)$ is applied to the bottom end of the column, then the capillary pressure head $h(z,t)$ can be shown to satisfy

$$\begin{aligned} &C(h)\partial_t h(z,t) = \partial_z(K(h)(\partial_z h(z,t) - 1)) && \text{for } 0 < z < L, \quad 0 < t < T, \\ &h(z,0) = z - 1 && \text{for } 0 < z < L, \\ &\partial_z h(0,t) - 1 = 0, \qquad h(L,t) = s(t) && \text{for } 0 < t < T. \end{aligned} \tag{3.1}$$

Here C and K continue to denote the water capacity and hydraulic conductivity, respectively. In fact, they are the same functions as seen previously since the phase-two experiment deals with the same soil subject, now for conditions designed to explore more of the range of the hydraulic functions C and K. The applied suction at the bottom, denoted by $s(t)$, is assumed to be smooth and monotone in time; i.e., $s(t)$ satisfies

$$s\varepsilon\mathbb{C}^1[0,T], \qquad s(0) = 0, \qquad s'(t) < 0 \quad \text{for } 0 < t < T. \tag{3.2}$$

Problem (3.1) is called the phase-two direct problem. For each pair of admissible coefficients (C, K), the direct problem (3.1) has a unique solution h whose dependence on the coefficients will be indicated by writing $h = \Psi_2[C, K]$. In the phase-two experiment, it is still relatively easy to measure the pressure head $h(0, t)$ at the top of the column and the flux or outflow, $K(h)(\partial_z h(1, t) - 1)$, at the bottom of the column. Likewise, for $h = \Psi_2[C, K]$, we can compute the functions

$$(3.3) \qquad p(t) = h(0, t) \quad \text{and} \quad q(t) = K(h)(\partial_z h(1, t) - 1) \quad \text{for } 0 < t < T.$$

The dependence of $p(t)$ and $q(t)$ on the coefficients C and K will be indicated by writing $(p, q) = \Gamma \cdot \Psi_2[C, K]$, and this association is called the *coefficient-to-data mapping* for the phase-two experiment. The functions (p, q) are viewed as output, and their properties are determined by and must be deduced from the properties of the input, particularly the suction $s(t)$. The following lemmas are the analogues of corresponding lemmas for the phase-one experiment and are needed in order to carry out the analysis of the inverse problem associated with the phase-two experiment.

LEMMA 3.1. *For admissible coefficients C and K, let $(p, q) = \Gamma \cdot \Psi[C, K]$. If $s(t)$ satisfies* (3.2), *then $q(t)$ defined in* (3.3) *satisfies*

$$(3.4) \qquad q \varepsilon \mathbb{C}[0, t], \qquad q(0) = 0, \quad \text{and} \quad q(t) < 0 \quad \text{for } 0 < t < T.$$

Proof. Let $\varphi(z, t) = s(t) + z - L$. Then $\partial_t \varphi(z, t) = s'(t)$ and $\partial_z \varphi(z, t) = 1$. Hence if $h(z, t)$ solves (3.1), it follows that $C(h)\partial_t(h - \varphi) = \partial_z(K(h)\partial_z(h - \varphi)) - C(h)s'(t)$. Then $w(z, t) = h(z, t) - \varphi(z, t)$ solves the initial boundary value problem

$$\begin{aligned} C(h)\partial_t w(z, t) - \partial_z(K(h)\partial_z w(z, t)) &= -C(h)s'(t) > 0, \\ w(z, 0) &= z - L - (z - L) = 0, \\ \partial_z w(0, t) &= \partial_z h(0, t) - \partial_z \varphi(0, t) = 0, \quad w(1, t) = s(t) - s(t) = 0. \end{aligned}$$

Then the maximum principle can be applied as in the proof of Lemma 1.1 and the conclusion follows.

LEMMA 3.2. *For admissible coefficients C and K, let $h = \Psi_2[C, K]$. If $s(t)$ satisfies* (3.2), *then $\partial_z h(z, t) - 1 < 0$ almost everywhere in Q_T.*

The proof of this and the next two lemmas are similar to the proofs of Lemmas 1.2, 1.3, and 1.4 and are omitted.

LEMMA 3.3. *For admissible coefficients C and K, let $h = \Psi_2[C, K]$. If $s(t)$ satisfies* (3.2), *then $\partial_t h(z, t) < 0$ almost everywhere in Q_T.*

LEMMA 3.4. *For admissible coefficients C and K, let $(p, q) = \Gamma \cdot \Psi_2[C, K]$. If $s(t)$ satisfies* (3.2), *then $h = \Psi_2[C, K]$ satisfies*

$$\textit{for each } \tau, 0 < \tau \le T, \quad s(\tau) + z - L < h(z, t) < z - L \quad \textit{for } 0 \le z \le L, \quad 0 \le t \le \tau$$

and $p(t) = h(0, t)$ satisfies

$$(3.5) \qquad p \varepsilon \mathbb{C}^1[0, T], \qquad p(0) = -L, \quad \textit{and} \quad p'(t) < 0 \quad \textit{for } 0 < t < T.$$

THEOREM 3.5. *For admissible coefficients C_j and K_j, $j = 1, 2$, let $h_j(z, t) = \Psi_2[C_j, K_j]$ and suppose $s(t)$ satisfies* (3.2). *Let $(p_j, q_j) = \Gamma \cdot \Psi_2[C_j, K_j]$, $j = 1, 2$. Then for any τ, $0 < \tau \le T$.*

$$(3.6) \quad \int_0^\tau \Delta p(t)\vartheta(t)\, dt = \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z, t) - 1)\partial_z \varphi_1 + \Delta C(h_2)\varphi_1 \partial_t h_2]\, dz\, dt,$$

$$(3.7) \quad \int_0^\tau \Delta q(t)\rho(t)\, dt = \iint_{Q_\tau} [\Delta K(h_2)(\partial_z h_2(z, t) - 1)\partial_z \varphi_2 + \Delta C(h_2)\varphi_2 \partial_t h_2]\, dz\, dt,$$

where $\varphi_1 = \Psi^*[\vartheta, 0]$ *and* $\varphi_2 = \Psi^*[0, \rho]$, *respectively.*

The proof of this theorem is nearly identical to the proof of Theorem 1.5 and is omitted.

Note that conditions (3.4) and (3.5) are necessary conditions for (p, q) to be a data pair generated from the phase-two experiment by admissible coefficients C and K. A function pair (p, q) as defined in (1.3) will be said to be a phase-two admissible data pair if they satisfy conditions (3.4) and (3.5), respectively.

4. The phase-two inverse problem. In the phase-one experiment the data $p(t)$ and $q(t)$ measured over the interval $[0, T]$ are used to determine the coefficients C and K on the interval $[-L, 0]$. In the phase-two experiment, the coefficients are assumed to be known on $[-L, 0]$, and a new experiment is conducted, applying suction to the column and again measuring $p(t)$ and $q(t)$ for a period of time. The time interval for the phase-two experiment will also be denoted by $[0, T]$, although T in phase two does not need to have a connection to the T for phase one.

In Lemma 3.1, it was shown that $h = \Psi_2[C, K]$ satisfies $h(z, t) > s(t) + z - L$ in Q_T. Then the domain of the coefficient-to-data mapping consists of pairs of functions that are continuous on the interval $[h_*, 0]$ for $h_* = -L + s(T)$. The range of the coefficient-to-data mapping consists of pairs of functions that are continuous on $[0, T]$. Thus while the phase-one inverse problem determines the coefficients C and K on the interval $[-L, 0]$, for $s(t)$ satisfying (3.2), the phase-two inverse problem will extend knowledge of the unknown coefficients to the interval $[-L + p(T), -L]$.

The next theorems are the phase-two analogues of Theorems 2.1, 2.3, and 2.4. The proofs of these theorems are virtually the same as the proofs of the phase-one theorems and are therefore omitted.

THEOREM 4.1. *For admissible coefficients* C_j *and* K_j, $j = 1, 2$, *let* $h_j(z, t) = \Psi_2[C_j, K_j]$ *and suppose* $s(t)$ *satisfies* (3.2). *Let* $(p_j, q_j) = \Gamma \cdot \Psi_2[C_j, K_j]$, $j = 1, 2$. *If* C_1, C_2 *and* K_1, K_2 *are distinguishable on the interval* $[h_*, 0]$, *then* p_1, p_2 *and* q_1, q_2 *are not identical on* $[0, T]$.

Theorem 4.1 asserts that distinguishable coefficients cannot produce phase-two experimental data pairs that are identical. Then the coefficient pairs are uniquely determined by the data collected for the phase-two experiment. The theorem may also be interpreted as asserting that the coefficient-to-data mapping $(p, q) = \Gamma \cdot \Psi_2[C, K]$ is an injection from a class of distinguishable functions into the set of admissible data. The injectivity of this mapping implies invertibility in the sense of Theorem 2.3.

THEOREM 4.2. *Let* $[p(t), q(t)]$ *denote a phase-two admissible data pair on* $[0, T]$, *and let* $\{0 = t_0 < \cdots < t_n = T\}$ *denote an arbitrary partition of the interval* $[0, T]$. *Suppose parameter pairs* $\{(\xi_0, \kappa_0), \ldots, (\xi_n, \kappa_n)\}$ *have been generated using the algorithm described in section* 2 *(modified to apply to the phase-two experiment) and that* $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ *for* $C^* = P_n[\xi_0, \ldots, \xi_n]$ *and* $K^* = P_n[\kappa_0, \ldots, \kappa_n]$. *Then*

(a) $\int_0^{t_m} (p(t) - p^*(t))\vartheta_m(t)\, dt = \int_0^{t_m} (q(t) - q^*(t))\rho_m(t)\, dt = 0, m = 1, \ldots, n;$

(b) *if* $p, q \varepsilon \mathbb{C}^2[t_{m-1}, t_m]$ *for* $m = 1, \ldots n$, *then* $|p(t_m) - p^*(t_m)| \leq \nu \Delta t^2$ *and* $|q(t_m) - q^*(t_m)| \leq \nu \Delta t^2$ *for* $m = 1, \ldots, n$ *for a positive constant* ν *depending on* p *and* q *and* $\Delta t = \max(t_m - t_{m-1})$.

Note that for each m, ξ_m, κ_m represent approximations to the values $C(-1 + s_m)$ and $K(-1 + s_m)$, respectively. Thus ξ_0 and κ_0 are the last values obtained in the phase-one experiment and are therefore known for the phase-two experiment.

THEOREM 4.3. *Let* $(p, q) = \Gamma \cdot \Psi_1[C, K]$ *for* (C, K) *an admissible coefficient pair, and let* $\{0 = t_0 < \cdots < t_n = T\}$ *denote an arbitrary partition of the interval* $[0, T]$. *Suppose that parameter pairs* $\{(\xi_0, \kappa_0), \ldots, (\xi_n, \kappa_n)\}$ *have been generated using the al-*

gorithm described previously and that $(p^*, q^*) = \Gamma \cdot \Psi_1[C^*, K^*]$ *for* $C^* = P_n[\xi_0, \ldots, \xi_n]$ *and* $K^* = P_n[\kappa_0, \ldots, \kappa_n]$*. If* $C, K \varepsilon \mathbb{C}^2[p_i, p_{i-1}]$ *for* $1 \leq i \leq n$*, then*

$$|\xi_i - C(p_i)| \leq \nu \Delta p \quad \text{and} \quad |\kappa_k - K(p_i)| \leq \nu \Delta p \quad \text{for } 1 \leq i \leq n$$

for a positive constant ν *depending on* C *and* K *and for* $\Delta p = \max\{p_{i-1} - p_i\}$.

Theorem 4.3 asserts that by applying suction to one end of a vertical column of soil, it is possible to extend the interval over which the hydraulic functions C and K can be determined. The extent of the domain of determination is then limited only by the experimental capabilities of the suction apparatus.

REFERENCES

[1] J. R. Cannon, *The One Dimensional Heat Equation*, Addison–Wesley, Reading, MA, 1984.

[2] J. R. Cannon and P. C. DuChateau, *Determination of unknown coefficients in parabolic operators from overspecified initial boundary data*, J. Heat Transfer, 100 (1978) pp. 503–507.

[3] J. R. Cannon and P. C. DuChateau, *Some asymptotic boundary behavior of solutions of nonlinear parabolic initial boundary value problem*, J. Math. Anal. Appl., 68 (1979), pp. 536–547.

[4] G. Chavent and P. Lemonnier, *Identification de la nonlinearite d'une equation parabolique quasilineare*, Appl. Math. Optim., 1 (1971), pp. 121–162.

[5] P. C. DuChateau, *Monotonicity and invertibility of coefficient to data mappings for parabolic inverse problems*, SIAM J. Math. Anal., 26 (1995), pp. 1473–1487.

[6] P. C. DuChateau, *Introduction to inverse problems in partial differential equations for engineers, scientists and engineers*, in Proc. Conference on Inverse Problems in Geology and Hydrology, Kluwer, Dordrecht, the Netherlands, 1996.

[7] B. H. Gilding, *Qualitative mathematical analysis of the Richard equation*, Transport Porous Media, 5 (1991), pp. 651–666.

[8] U. Hornung, *Identification of nonlinear soil parameters from an input-output experiment*, in Progress in Scientific Computing, Birkhäuser Boston, Cambridge, MA, 1983.

[9] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monographs 23, AMS, Providence, RI, 1969.

[10] N. V. Muzylev, *Uniqueness theorems for some converse problems in heat conduction*, USSR Comput. Math. Math. Phys., 20 (1986), pp. 120–134.

[11] D. Zachmann, P. DuChateau, and A. Klute, *Simultaneous approximation of water capacity and soil hydraulic conductivity by parameter identification*, Soil Sci., 134 (1982), pp. 157–163.

[12] J. R. Cannon, P. DuChateau, and K. Steube, *Trace type functional differential equations and the identification of hydraulic properties of porous media*, Transport Porous Media, 6 (1991), pp. 745–758.

SIAM J. Math. Anal.
Vol. 28, No. 3, pp. 633–655, May 1997

ON CONCENTRATION OF POSITIVE BOUND STATES OF NONLINEAR SCHRÖDINGER EQUATIONS WITH COMPETING POTENTIAL FUNCTIONS[*]

XUEFENG WANG[†] AND BIN ZENG[†]

Abstract. Some nonlinear Schrödinger equations with several competing potential functions are considered. Ground states (least energy solutions) are proved to exist and concentrate at a point in the semiclassical limit. The concentration points are shown to be located on the middle ground of the competing potential functions and in some cases are given explicitly in terms of these functions. Also given is a necessary condition for location of concentration of positive bound states (solutions with higher but finite energy).

Key words. nonlinear Schrödinger equations, competing potentials, ground states, bound states, energy, semiclassical limit, concentration, existence

AMS subject classifications. 81Q05, 35J60, 35B25

PII. S0036141095290240

1. Introduction.

The simplest equation considered in this paper is the following:

$$h^2\Delta u - V(x)u + K(x)|u|^{p-1}u = 0, \quad x \in R^n, \tag{1}$$

where $1 < p < (n+2)/(n-2)^+$ ($= \infty$ if $n = 1, 2$) and $V(x)$ and $K(x)$ are positive smooth functions with $V(x)$ bounded below by a positive constant and $K(x)$ bounded. We are interested in the existence and concentration behavior of positive *ground states* of (1) and its generalization (4) (see below) in the semiclassical limit $h \to 0$. We are also interested in finding a necessary condition for location of concentration of positive *bound states* of these equations. A ground state of (1) is a solution of the equation which has the least energy,

$$\frac{1}{2}\int_{R^n}(h^2|\nabla u|^2 + V(x)u^2)\,dx - \frac{1}{p+1}\int_{R^n}K(x)|u|^{p+1}\,dx,$$

among all nontrivial solutions of (1). A bound state is a solution with finite but not necessarily least energy.

Equation (1) arises at least in the study of standing-wave solutions of the following time-dependent nonlinear Schrödinger equation:

$$ih\frac{\partial\psi}{\partial t} = -h^2\Delta\psi + V(x)\psi - K(x)|\psi|^{p-1}\psi, \quad x \in R^n.$$

It also provides standing-wave solutions of the Klein–Gorden equations. See [1] and the references therein for more background. Many authors [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [13], [16], [17], [18], [19], [20] have worked on equation (1) in various forms and obtained numerous results on existence, uniqueness, and radial symmetry

[*] Received by the editors August 11, 1995; accepted for publication January 25, 1996.
http://www.siam.org/journals/sima/28-3/29024.html

[†] Department of Mathematics, Tulane University, New Orleans, LA 70118 (xdw@math.tulane.edu). The research of the first author was partially supported by NSF grant DMS-9305658.

of solutions. Of particular interest in this paper is the following existence result of Rabinowitz [19]: if

$$K(x) \equiv 1 \quad \text{and} \quad \liminf_{|x|\to\infty} V(x) > \inf_{x\in R^n} V(x) > 0, \tag{2}$$

then for small h, (1) has a positive ground state. Subsequently, Wang [22] addressed the concentration problem of bound states of (1). For ground states, he proved that under condition (2), any sequence of ground states contains a subsequence which concentrates at a global minimum of $V(x)$ as h tends to zero. Moreover, if we hold $V(x)$ constant instead of $K(x)$, by essentially the same proof as in [22], we can show that ground states concentrate at a global maximum of $K(x)$. When neither $V(x)$ nor $K(x)$ is constant, there would presumably be competition between $V(x)$ and $K(x)$: each would try to attract ground states to their minimum and maximum points, respectively. Now some natural questions arise: Do ground states still concentrate? If so, where? These questions are the primary motivation of this paper.

By variational techniques as developed by Rabinowitz [19] and Wang [22], we show that under certain conditions on the potential functions $V(x)$ and $K(x)$, positive ground states still exist for small h and as h approaches zero, they concentrate at a point on the "middle ground" between "valleys" of $V(x)$ and "peaks" of $K(x)$. More precisely, we prove that if

$$\frac{\liminf_{|x|\to\infty} V^{(2p+2+n-np)/(2p-2)}(x)}{\limsup_{|x|\to\infty} K^{2/(p-1)}(x)} > \inf_{x\in R^n} \frac{V^{(2p+2+n-np)/(2p-2)}(x)}{K^{2/(p-1)}(x)}, \tag{3}$$

then positive ground state solutions of (1) *exist* for small positive h and *concentrate* at a global minimum point of $g(x) := V^{(2p+2+n-np)/(2p-2)}(x)/K^{2/(p-1)}(x)$ as h approaches zero (see Corollaries 2.9 and 3.2). We remark that if (3) does not hold, even the existence is in question. However, if we know that a sequence $\{u_{h_k}\}$ of positive ground states of (1) exists with each having a local maximum point moving toward a point x_0 as $h_k \to 0$, then x_0 is a global minimum point of $g(x)$ and $\{u_{h_k}\}$ concentrates at x_0. Consequently, in the even worse scenario in which $g(x)$ does not have a minimum point, the positive ground states, if any, do not concentrate as h shrinks (they move off to infinity).

To gain further insight into the effect of potential functions on the concentration process, we also consider the following generalization of (1):

$$h^2\Delta u - V(x)u + K(x)|u|^{p-1}u + Q(x)|u|^{q-1}u = 0, \quad x \in R^n, \tag{4}$$

where $1 < q < p < (n+2)/(n-2)^+$ and $Q(x)$ is a bounded smooth function that is allowed to change its sign. When $V(\infty) = \infty$, a result in [19] implies the existence of ground states of (4) for any $h > 0$. In other cases, we cannot find any result in the literature that can be directly applied to obtain even the existence of ground states. The *ground-energy function* $C(s)$, which is defined to be the least (or ground) energy associated with

$$\Delta u - V(s)u + K(s)|u|^{p-1}u + Q(s)|u|^{q-1}u = 0, \quad x \in R^n, \tag{5}$$

where $s \in R^n$ is regarded as a parameter instead of an independent variable, plays a central role in our results and analysis. (See section 2 for the precise definition.) Let c_∞ be the ground energy associated with

$$\Delta u - V_\infty u + K_\infty |u|^{p-1}u + Q_\infty |u|^{q-1}u = 0, \quad x \in R^n,$$

where

$$V_\infty = \liminf_{|s|\to\infty} V(s), \qquad K_\infty = \limsup_{|s|\to\infty} K(s), \qquad Q_\infty = \limsup_{|s|\to\infty} Q(s).$$

We prove that if

$$c_\infty > \inf_{s\in R^n} C\,(s),$$

then for small h (4) has a positive ground state u_h and as h shrinks to zero, after passing to a subsequence u_h concentrates at a global minimum point of $C(s)$ (see Theorems 2.7 and 3.1). From this we will obtain some concrete sufficient conditions (explicitly expressed in terms of the potential functions) for the existence of positive ground states of (4) with small h—see Corollary 2.8. As will be seen, in the case of (1), $g(s)$ is just a positive constant multiple of the ground-energy function $C(s)$ (see (34)). Thus the results for (1) described above are consequences of these general results for (4). However, unlike (1), in the general case (4), we do not (and it is impossible to) have an explicit formula for $C(s)$, and hence we cannot explicitly express the location of concentration of ground states in terms of the potential functions $V(x)$, $K(x)$, and $Q(x)$.

In this paper, we also provide a necessary condition for location of concentration of positive bound states (not necessarily with least energy) of (4). We prove that a point x_0 at which a sequence of positive bound states concentrates as $h \to 0$ must be a critical point of the ground-energy function $C(s)$, provided that (5) with $s = x_0$ has at most one positive decaying solution (up to translations); in the special case (1), this means that the concentration points of positive bound states can only be critical points of $g(x) := V^{(2p+2+n-np)/(2p-2)}(x)/K^{2/(p-1)}(x)$. (See Theorem 4.1 for the precise statement as well as some specific conditions that ensure the uniqueness of aforementioned solutions of (5).) This generalizes the corresponding result in Wang [22], where the case $K(x) \equiv 1$ is considered. This is also in the converse direction of the following result of Floer, Weinstein, and Oh [6], [16], [17]: for small h, (1) with $K(x) \equiv 1$ has a positive "multibump" bound state which concentrates at any given finite set of nondegenerate critical points of $V(x)$ under the condition, say, that $V(x)$ is bounded. (See [22] for a remark on this condition.)

We mention that in Zeng's thesis [24], for equation (1), an existence result in the spirit of [6], [16], and [17] is established under a technical nondegeneracy condition. In particular, if both $V(x)$ and $K(x)$ share a common critical point x_0 (so x_0 is a critical point of $g(x)$) and if none of $(\partial^2/\partial x_i^2)g\,(x_0)$, $i = 1, 2, \ldots, n$, is zero, then for small h, (1) has a positive bound state concentrating at x_0 as $h \to 0$. In this connection, after our work was completed, we received an interesting preprint by Gui, "Existence of multi-bump solutions for nonlinear Schrödinger equations via variational method," which deals with the equation

$$h^2\Delta u - V(x)u + f(u) = 0, \quad x \in R^n. \tag{6}$$

Gui proves the existence of bound states concentrating at finitely many local (not necessarily strict) minimum points of $V(x)$. However, he does not consider the competing potential $K(x)$. Rabinowitz later informed us that, in his thesis, Thandi proved the existence of solutions of (6) with infinitely many bumps and that Del Pino and Felmer also obtained existence results similar to Gui's independently.

We remind the reader that *throughout this paper, we always make the following assumptions*:

(H_1): $V(x)$, $K(x)$, and $Q(x)$ are C^1 smooth in R^n.

(H_2): $\inf_{x\in R^n} V(x) = \overline{V} > 0$, and both $K(x)$ and $Q(x)$ are bounded in R^n, with $K(x) > 0$ and $Q(x)$ allowed to change sign.

2. Existence of ground states.

2.1. Preliminaries. We start by transforming equation (4). Let $v(x) = u(hx)$. Then equation (4) becomes

$$\Delta v(x) - V(hx)v + K(hx)|v|^{p-1}v + Q(hx)|v|^{q-1}v = 0, \quad x \in R^n. \tag{7}$$

Since equations (4) and (7) are equivalent, we shall thereafter focus on equation (7).

Let E_h be the Hilbert subspace of $v \in H^1(R^n)$ under the norm

$$\|v\|^2_{E_h} := \int_{R^n} (|\nabla v|^2 + V(hx)v^2)\, dx < +\infty.$$

Define the *energy functional* associated with (7) by

$$\begin{aligned} I_h(v) := \frac{1}{2}\int_{R^n} (|\nabla v|^2 + V(hx)v^2)\, dx - \frac{1}{p+1}\int_{R^n} K(hx)|v|^{p+1}\, dx \\ - \frac{1}{q+1}\int_{R^n} Q(hx)|v|^{q+1}\, dx \end{aligned} \tag{8}$$

for $v \in E_h$ and $h > 0$, and define the *solution manifold* of (7) by

$$\begin{aligned} M_h := \Big\{ v \in E_h\backslash\{0\} : \int_{R^n} (|\nabla v|^2 + V(hx)v^2)\, dx = \int_{R^n} K(hx)|v|^{p+1}\, dx \\ + \int_{R^n} Q(hx)|v|^{q+1}\, dx \Big\}. \end{aligned} \tag{9}$$

(Any nontrivial solution of (7) in $H^1(R^n)$ belongs to M_h; hence we have the name for M_h.) A *ground state* of (7) is defined as a solution of (7) which minimizes $I_h(v)$ on M_h. The *ground energy* associated with (7) is defined as

$$c_h := \inf_{v\in M_h} I_h(v). \tag{10}$$

By the Sobolev imbedding theorem, it is easy to see that the ground energy c_h is finite. It is also routine to show that (i) any minimizer of I_h on M_h is a solution of (7); (ii) any minimizer is of one sign and since I_h is even, we can always take it to be positive. Thus to show that (7) has a positive ground state, we only need to show that I_h has a minimizer over M_h.

We now show that these minimizers can be found through a minimax process. Let

$$\Gamma_h := \{\, \eta \in C([0,1], E_h) : \eta(0) = 0,\ \eta(1) \not\equiv 0,\ I_h(\eta(1)) \le 0\,\} \tag{11}$$

and define

$$c^*_h := \inf_{\eta\in\Gamma_h} \max_{0\le t\le 1} I_h(\eta(t)). \tag{12}$$

Thus c^*_h is the mountain-pass minimax value associated with I_h.

Define yet another minimax value,

$$c_h^{**} := \inf_{v\in E_h\backslash\{0\}} \max_{t\geq 0} I_h(tv). \tag{13}$$

Lemma 2.1. $c_h = c_h^* = c_h^{**}$.

Remark. This kind of result has already been observed [14], [19] for equations which do not include our (7) or (4) when Q is negative. From this result, we see that to show the existence of ground states of (7), we just need to prove that the minimax value given by (12) and (13) is a critical value of I_h, though we shall not strictly follow this course when proving existence in the future.

Proof. We first show that $c_h = c_h^{**}$. This will follow if we can show that for any $v \in E_h\backslash\{0\}$, the ray $R_t = \{tv : t \geq 0\}$ intersects the solution manifold M_h once and only once at $\theta v\,(\theta > 0)$, where $I_h(tv)$, $t \geq 0$, achieves its maximum. A direct computation shows that critical points of the function $f(t) := I_h(tv)$ occur at and only at the intersections of the ray R_t and M_h. It is easy to see that the ray R_t intersects M_h. Suppose this occurs first at $t = t_0 > 0$. Let

$$g(t) = At^{p-1} + Bt^{q-1}, \tag{14}$$

where $A = \int_{R^n} K(hx)|v|^{p+1}\,dx$ and $B = \int_{R^n} Q(hx)|v|^{q+1}\,dx$. Then

$$\int_{R^n} (|\nabla v|^2 + V(hx)v^2)\,dx = g(t_0).$$

On the other hand, it is elementary to show that

$$g(t) \text{ is strictly increasing on any interval where } g(t) > 0. \tag{15}$$

Thus $g(t)$ is strictly increasing for $t \geq t_0$, and hence the ray R_t intersects M_h only once. We have shown that $c_h = c_h^{**}$.

It remains to show that $c_h = c_h^*$. The inequality $c_h^* \leq c_h^{**} = c_h$ is true because for each nonzero $v \in E_h$, there exists a segment of the half-line $\{tv : t \geq 0\}$ that contains the maximum point of $I_h(tv)$ and is a path in Γ_h. Now we show that $c_h \leq c_h^*$.

By the Sobolev embedding theorem, for nonzero v with $\|v\|_{E_h}$ small,

$$\int_{R^n} (|\nabla v|^2 + V(hx)v^2)\,dx > \int_{R^n} K(hx)|v|^{p+1}\,dx + \int_{R^n} Q(hx)|v|^{q+1}\,dx. \tag{16}$$

We claim that any path $\eta(t)$ in Γ_h crosses M_h. Otherwise, by the continuity of $\eta(t)$, inequality (16) still holds true when v is replaced by each nonzero $\eta(t)$ for $t \in [0,1]$. Recall that $\eta(1)$ is nonzero. Then

$$\begin{aligned} I_h(\eta(1)) &= \frac{1}{2}\int_{R^n} (|\nabla\eta|^2 + V(hx)\eta^2)\,dx - \frac{1}{p+1}\int_{R^n} K(hx)|\eta|^{p+1}\,dx \\ &\quad - \frac{1}{q+1}\int_{R^n} Q(hx)|\eta|^{q+1}\,dx \\ &> \frac{1}{q+1}\int_{R^n} (|\nabla\eta|^2 + V(hx)\eta^2)\,dx - \frac{1}{p+1}\int_{R^n} K(hx)|\eta|^{p+1}\,dx \\ &\quad - \frac{1}{q+1}\int_{R^n} Q(hx)|\eta|^{q+1}\,dx \\ &\geq \left(\frac{1}{q+1} - \frac{1}{p+1}\right)\int_{R^n} K(hx)|\eta|^{p+1}\,dx \\ &> 0. \end{aligned}$$

The above inequality violates the definition of $\eta(1)$. Thus $\eta(t)$ crosses M_h, and hence

$$\max_{0\le t\le 1} I_h(\eta(t)) \ge \inf_{v\in M_h} I_h(v) = c_h.$$

This completes the proof of Lemma 2.1. ∎

Now we discuss the properties of the ground energy c_h given by (10). Denote c_1 by $c(V,K,Q)$. The following comparison result is similar to the one in [19].

LEMMA 2.2. *Suppose that $V_a(x)$, $V_b(x)$, $K_a(x)$, $K_b(x)$, $Q_a(x)$, and $Q_b(x)$ satisfy our* (*standing*) *conditions* (H_1) *and* (H_2). *If*

$$V_a \le V_b, \qquad K_a \ge K_b, \qquad Q_a \ge Q_b, \tag{17}$$

then

$$c\,(V_a, K_a, Q_a) \le c\,(V_b, K_b, Q_b).$$

If, in addition, one of the inequalities in (17) *is strict and $V_b(x)$, $K_b(x)$, and $Q_b(x)$ are constant functions, then*

$$c\,(V_a, K_a, Q_a) < c\,(V_b, K_b, Q_b).$$

Proof. Let I^a be the energy functional associated with $c\,(V_a,K_a,Q_a)$. Define other related notation in the obvious way. Note that $E^b \subset E^a$ and for any $v \in E^b$, $I^a(v) \le I^b(v)$. Thus by Lemma 2.1,

$$\begin{aligned} c\,(V_b, K_b, Q_b) &= \inf_{v\in E^b\setminus\{0\}} \max_{t\ge 0} I^b(tv) \ge \inf_{v\in E^a\setminus\{0\}} \max_{t\ge 0} I^a(tv) \\ &= c\,(V_a, K_a, Q_a). \end{aligned}$$

Now we prove the second assertion. Since $V_b(x)$, $K_b(x)$, and $Q_b(x)$ are constants, we have that $E^b = E^a = H^1$. Furthermore, it is well known that there exists a ground state $v_b \in H^1$ such that $c\,(V_b,K_b,Q_b) = I^b(v_b)$. (See, e.g., [19, Theorem 4.23]. We should mention that condition (f_5) in [19] is not satisfied if $Q_b < 0$, but that condition is used there to show identities like those in our Lemma 2.1.) Now we have

$$\begin{aligned} c\,(V_b, K_b, Q_b) = I^b(v_b) \ &= \ \max_{t\ge 0} I^b(tv_b) \ > \ \max_{t\ge 0} I^a(tv_b) \\ &\ge \inf_{v\in H^1\setminus\{0\}} \max_{t\ge 0} I^a(tv) \ = \ c\,(V_a, K_a, Q_a). \end{aligned}$$

Here we used Lemma 2.1. This proves Lemma 2.2. ∎

Next, we define the *ground-energy function* $C(s)$ mentioned in section 1. For each $s \in R^n$, consider

$$\Delta v(x) - V(s)v + K(s)|v|^{p-1}v + Q(s)|v|^{q-1}v = 0, \quad x \in R^n. \tag{18}$$

Let I^s be the associated energy functional, i.e.,

$$\begin{aligned} I^s(v) := \frac{1}{2}\int_{R^n} (|\nabla v|^2 + V(s)v^2)\,dx - \frac{1}{p+1}\int_{R^n} K(s)|v|^{p+1}\,dx \\ - \frac{1}{q+1}\int_{R^n} Q(s)|v|^{q+1}\,dx. \end{aligned} \tag{19}$$

Define the *ground-energy function* by

$$C\,(s) = \inf_{v \in M^s} I^s(v), \tag{20}$$

where M^s is the solution manifold of (18) defined as in (9). As mentioned in the proof of Lemma 2.1, it is well known that for each $s \in R^n$, (18) has a positive ground state $v_s(x)$ (i.e., $v_s(x)$ solves (20)). By [7], v_s is spherically symmetric about, say, the origin and is decreasing in $r = |x|$.

LEMMA 2.3.

(i) *The ground energy function $C(s)$ is locally Lipschitz continuous in $s \in R^n$.*

(ii) *Denote by $C_i^l(s)$ and $C_i^r(s)$, $i = 1, 2, 3, \ldots, n$, the left and right partial derivatives of $C(s)$ with respect to ith variable s_i, respectively. Then they always exist at all $s \in R^n$ and*

$$\begin{aligned} C_i^l(s) = \sup_{v_s \in G^s} \bigg[&\frac{1}{2}\frac{\partial V(s)}{\partial s_i} \int_{R^n} v_s^2(x)\,dx \\ &- \frac{1}{p+1}\frac{\partial K(s)}{\partial s_i} \int_{R^n} v_s^{p+1}(x)\,dx - \frac{1}{q+1}\frac{\partial Q(s)}{\partial s_i} \int_{R^n} v_s^{q+1}(x)\,dx \bigg], \end{aligned} \tag{21}$$

$$\begin{aligned} C_i^r(s) = \inf_{v_s \in G^s} \bigg[&\frac{1}{2}\frac{\partial V(s)}{\partial s_i} \int_{R^n} v_s^2(x)\,dx \\ &- \frac{1}{p+1}\frac{\partial K(s)}{\partial s_i} \int_{R^n} v_s^{p+1}(x)\,dx - \frac{1}{q+1}\frac{\partial Q(s)}{\partial s_i} \int_{R^n} v_s^{q+1}(x)\,dx \bigg], \end{aligned} \tag{22}$$

where $i = 1, 2, 3, \ldots, n$ and $G^s =$ the set of all positive (radial) ground states of (18).

(iii) *The function $C(s)$ is C^1 smooth throughout R^n in any of the following cases:* (a) $n = 1$; (b) $Q(s) \le 0$ *in* R^n; (c) $1 < q < p \le n/(n-2)$ *and* $n > 2$.

Remark. In general, by Rademacher's theorem and (i), $C(s)$ is differentiable almost everywhere in R^n. Also, if (18) has a unique positive ground state (up to translations) for each s, then $C(s)$ is C^1 smooth. Unfortunately, the uniqueness is known only in the cases listed in (iii). We point out that if (b) in part (iii) holds at a point, then all partial derivatives of $C(s)$ exist at that point.

Proof. As discussed in the proof of Lemma 2.1, for any s, $t \in R^n$, there is a unique positive constant $\theta(s,t)$ such that $\theta(s,t)v_t(x) \in M^s$. Thus

$$\begin{aligned} &\int_{R^n} (|\nabla v_t(x)|^2 + V(s)v_t^2(x))\,dx \\ &\quad = \theta^{p-1}(s,t) \int_{R^n} K(s)v_t^{p+1}(x)\,dx + \theta^{q-1}(s,t) \int_{R^n} Q(s)v_t^{q+1}(x)\,dx \end{aligned} \tag{23}$$

and $\theta(t,t) = 1$. By the implicit-function theorem, $\theta(s,t)$ is differentiable with regard to variable s.

CLAIM 2.3.1. *$\theta(s,t)$ is bounded for bounded s and t.*

To see this, we first observe that

$$\begin{aligned} C(t) = I^t(v_t) &= \left(\frac{1}{2} - \frac{1}{q+1}\right) \int_{R^n} (|\nabla v_t(x)|^2 + V(t)v_t^2(x))\,dx \\ &\quad + \left(\frac{1}{q+1} - \frac{1}{p+1}\right) \int_{R^n} K(t)v_t^{p+1}(x)\,dx \\ &= \left(\frac{1}{2} - \frac{1}{p+1}\right) \int_{R^n} K(t)v_t^{p+1}(x)\,dx \end{aligned}$$

$$+\left(\frac{1}{2}-\frac{1}{q+1}\right)\int_{R^n} Q(t)v_t^{q+1}(x)\,dx. \tag{24}$$

From this, Lemma 2.2, and the Sobolev imbedding theorem, we have that $C(t)$ and the H^1, L^{p+1}, and L^{q+1} norms of v_t are bounded for bounded t; moreover, for such a t, $C(t)$, and the L^{p+1} norm of v_t are bounded from below by a positive constant. Now from (23), we see that Claim 2.3.1 is true.

To prove (i), we first note that the gradient with respect to variable s of

$$\begin{aligned}
&I^s(\theta(s,t)v_t(x))\\
&\quad=\frac{1}{2}\theta^2(s,t)\int_{R^n}(|\nabla v_t(x)|^2+V(s)v_t^2(x))\,dx\\
&\qquad-\frac{1}{p+1}\theta^{p+1}(s,t)\int_{R^n}K(s)v_t^{p+1}(x)\,dx-\frac{1}{q+1}\theta^{q+1}(s,t)\int_{R^n}Q(s)v_t^{q+1}(x)\,dx.
\end{aligned}$$

is given by

$$\begin{aligned}
&\nabla_s I^s(\theta(s,t)v_t)\\
&\quad=\frac{1}{2}\theta^2(s,t)\nabla_s V(s)\int_{R^n}v_t^2(x)\,dx-\frac{1}{p+1}\theta^{p+1}(s,t)\nabla_s K(s)\int_{R^n}v_t^{p+1}(x)\,dx\\
&\qquad-\frac{1}{q+1}\theta^{q+1}(s,t)\nabla_s Q(s)\int_{R^n}v_t^{q+1}(x)\,dx\\
&\qquad+\nabla_s\theta(s,t)\bigg[\theta^2(s,t)\int_{R^n}(|\nabla v_t(x)|^2+V(s)v_t^2(x))\,dx\\
&\qquad\quad-\theta^{p+1}(s,t)\int_{R^n}K(s)v_t^{p+1}(x)\,dx-\theta^{q+1}(s,t)\int_{R^n}Q(s)v_t^{q+1}(x)\,dx\bigg]\Big/\theta(s,t)\\
&\quad=\frac{1}{2}\theta^2(s,t)\nabla_s V(s)\int_{R^n}v_t^2(x)\,dx-\frac{1}{p+1}\theta^{p+1}(s,t)\nabla_s K(s)\int_{R^n}v_t^{p+1}(x)\,dx\\
&\qquad-\frac{1}{q+1}\theta^{q+1}(s,t)\nabla_s Q(s)\int_{R^n}v_t^{q+1}(x)\,dx
\end{aligned}\tag{25}$$

because $\theta(s,t)v_t(x)\in M^s$. Thus by Claim 2.3.1, for any $R>0$, there exists a constant M such that

$$|\nabla_s I^s(\theta(s,t)v_t)|\le M,\quad |s|,\,|t|\le R. \tag{26}$$

From this it follows that for any $|s^1|$, $|s^2|\le R$,

$$\begin{aligned}
C(s^1)-C(s^2)&\le I^{s^1}(\theta(s^1,\,s^2)v_{s^2})-I^{s^2}(v_{s^2})\\
&=(s^1-s^2)\cdot\nabla_s I^s(\theta(s,s^2)v_{s^2}(x))\big|_{s=\xi\in[s^1,\,s^2]}\\
&\le M|s^1-s^2|,
\end{aligned}$$

where $[s^1,\,s^2]$ is the segment connecting s^1 and s^2.

Similarly, we can show

$$C(s^1)-C(s^2)\ge -M|s^1-s^2|.$$

This finishes the proof of part (i) of Lemma 2.3.

Claim 2.3.2. *For any sequence $\{s_k\}\to s_0$, there exists a subsequence* (*still denoted by $\{s_k\}$*) *such that $v_{s_k}\to w_0$ strongly in H^1, where w_0 is a positive ground state of* (18) *with $s=s_0$.*

As noted earlier, $\|v_s\|_{H^1}$ is bounded for bounded s. Therefore, by the standard elliptic regularity theory, we have, after passing to a subsequence,

$$v_{s_k} \to \text{some } w_0 \quad \text{weakly in } H^1 \text{ and strongly in } C^2_{\text{loc}}, \tag{27}$$

where $w_0 \geq 0$ satisfies (18) with $s = s_0$.

Since

$$-V(s_k)v_{s_k}(0) + K(s_k)|v_{s_k}|^{p-1}v_{s_k}(0) + Q(s_k)|v_{s_k}|^{q-1}v_{s_k}(0) = -\Delta v_{s_k}(0) \geq 0$$

(recall that $v_s(x)$ is radially decreasing), we have

$$K(s_0)|w_0|^{p-1}(0) + Q(s_0)|w_0|^{q-1}(0) \geq V(s_0) > 0.$$

Thus $w_0 \not\equiv 0$, and hence by the strong-maximum principle, $w_0(x) > 0$. Since it belongs to H^1 space, it is easy to show that $w_0(x) \to 0$ as $x \to \infty$. From this, (27), and the fact that $v_s(x)$ is radially decreasing, we obtain that for any small $\delta > 0$, there exists a constant $\rho > 0$ such that

$$|v_{s_k}(x)| < \delta \quad \text{for } |x| > \rho \quad \text{and} \quad k = 1, 2, \ldots. \tag{28}$$

Now take a sufficiently small δ such that

$$\overline{V} - \|K\|_{L^\infty}\, \delta^{p-1} - \|Q\|_{L^\infty}\, \delta^{q-1} > \frac{1}{2}\overline{V},$$

where $\overline{V} = \inf V(x)$. Then by (18) and (28), v_{s_k} is a subsolution of

$$\Delta w - \frac{1}{2}\overline{V}w = 0 \quad \text{for } |x| > \rho.$$

By the comparison principle,

$$|v_{s_k}(x)| \leq \delta \exp\{-C_1(|x| - \rho)\} \quad \text{for } |x| \geq \rho, \tag{29}$$

where C_1 is a positive constant independent of k. Moreover, by virtue of the elliptic interior estimates, the gradient of $v_{s_k}(x)$ also decays exponentially fast at ∞ uniformly with respect to k. Combining these estimates for $v_{s_k}(x)$ with (27), we deduce that $(\rho' > \rho)$

$$\|v_{s_k} - w_0\|^2_{H^1(R^n)} = \|v_{s_k} - w_0\|^2_{H^1(B_{\rho'})} + \int_{B^c_{\rho'}} (|\nabla v_{s_k} - \nabla w_0|^2 + |v_{s_k} - w_0|^2)\, dx$$

converges to zero as $k \to \infty$. From this and the continuity of $C(s)$ (already proved), it follows that w_0 is a ground state of (18) with $s = s_0$. This completes the proof of Claim 2.3.2.

Now we proceed to prove part (ii). To save notation, we assume without loss of generality that $s \in R^1$, i.e., $n = 1$. For the same reason, we shall prove (21) and (22) only at $s = 0$.

For any $v_0 \in G^0$, the set of all positive (radial) ground states of (18) with $s = 0$, and for any $r > 0$, we have

$$\begin{aligned} C(r) - C(0) &\leq I^r(\theta(r,0)v_0) \; - \; I^0(v_0) \\ &= r\nabla_s\, I^s(\theta(s,0)v_0)\big|_{s=\xi\in[0,\,r]}. \end{aligned}$$

This, (25), and the fact that $\theta(s,0) \to \theta(0,0) = 1$ as $s \to 0$ imply that

$$\limsup_{r\to 0^+} \frac{C(r)-C(0)}{r} \le \inf_{v_0 \in G^0} \left[\frac{1}{2}V'(0)\int v_0^2(x)\,dx \right.$$

$$\left. - \frac{1}{p+1}K'(0)\int v_0^{p+1}(x)\,dx - \frac{1}{q+1}Q'(0)\int v_0^{q+1}(x)\,dx\right]. \tag{30}$$

On the other hand,

$$\begin{aligned} C(r) - C(0) &\ge I^r(\theta(r,r)v_r) \;-\; I^0(\theta(0,r)v_r) \\ &= r\nabla_s\, I^s(\theta(s,r)v_r)\big|_{s=\xi\in[0,\,r]}. \end{aligned}$$

From this, Claim 2.3.2, and (25), we have

$$\liminf_{r\to 0^+} \frac{C(r)-C(0)}{r} \ge \text{ right-hand side of (30).}$$

Thus (22) is true. (21) follows similarly.

To prove part (iii), we notice that in all cases (a)–(c), we have the uniqueness (up to translations) of positive ground states of (18) (see [1] when $n = 1$ and [2], [9], and [15, Appendix C] in cases (b) and (c)), and consequently the right and left partial derivatives of $C(s)$ are equal at all s. Furthermore, by Claim 2.3.2, the partial derivatives given by (21) or (22) are continuous in s. This completes the proof of Lemma 2.3. □

Using the arguments in the proof of Lemma 2.3(i), we easily see that the following is true.

LEMMA 2.4. *If V, K, and Q are constant functions, $c(V, K, Q)$ (defined before the statement of Lemma 2.2) depends continuously on them.*

We now present an expression for $C(s)$ which, in particular, enables us to express $C(s)$ explicitly in terms of $V(s)$ and $K(s)$ in the case where $Q(s) = 0$. Substituting $v(x) = \lambda w(\mu x)$ with $\mu^2 = V(s)$, $\lambda = [V(s)/K(s)]^{1/(p-1)}$, and

$$\alpha(s) = \frac{Q(s)}{V^{\frac{p-q}{p-1}}(s)\,K^{\frac{q-1}{p-1}}(s)}, \tag{31}$$

into (18), we have

$$\Delta w(x) - w + |w|^{p-1}w + \alpha(s)|w|^{q-1}w = 0. \tag{32}$$

LEMMA 2.5. *Let $c_{\alpha(s)} := c\,(1, 1, \alpha(s))$, i.e., the ground energy associated with (32). Then $c_{\alpha(s)}$ is a decreasing function of $\alpha(s)$ and*

$$C\,(s) = \frac{V^{\frac{p+1}{p-1}-\frac{n}{2}}(s)}{K^{\frac{p+1}{p-1}-1}(s)}c_{\alpha(s)}. \tag{33}$$

In particular, if $Q(s) = 0$, then

$$C\,(s) = \left(\frac{1}{2} - \frac{1}{p+1}\right)\frac{V^{\frac{p+1}{p-1}-\frac{n}{2}}(s)}{K^{\frac{p+1}{p-1}-1}(s)}\int_{R^n} U^{p+1}(x)\,dx, \tag{34}$$

where $U(x)$ is the unique positive ground state (up to translation) of (32) with $\alpha = 0$.

Proof. The monotone property of $c_{\alpha(s)}$ follows from Lemma 2.2.

Since (18) and (32) are related as indicated above, by direct computations, we obtain (33) and (34). □

We close this subsection by a result which will be very important to the existence and concentration results.

LEMMA 2.6. *There exists positive constant $\bar{c}$ such that $c_h > \bar{c}$. On the other hand,*

$$\limsup_{h \to 0^+} c_h \le \inf_{s \in R^n} C(s).$$

Proof. By Lemma 2.2, $c_h \ge c(\inf V, \|K\|_{L^\infty}, \|Q\|_{L^\infty})$, which is positive (as can be seen from an expression like (24)).

The remaining part of lemma can be proved by choosing good test functions. See the proof of Lemma 2.2 in [22] for a similar situation. □

2.2. Existence results. Define

$$V_\infty = \liminf_{|s| \to \infty} V(s), \qquad K_\infty = \limsup_{|s| \to \infty} K(s), \qquad Q_\infty = \limsup_{|s| \to \infty} Q(s). \tag{35}$$

Let $c_\infty := c\,(V_\infty, K_\infty, Q_\infty)$. If $V_\infty = +\infty$, define $c_\infty := +\infty$ (see (33)).

THEOREM 2.7. *If*

$$c_\infty > \inf_{s \in R^n} C\,(s), \tag{36}$$

then for small $h > 0$, equation (7) (*and hence* (4)) *has a positive ground-state solution.*

Proof. By Lemma 2.1, we can find a sequence $\{u_m\}$ such that $\|u_m\|_{E_h} = 1$ and

$$\max_{t \ge 0} I_h(tu_m) \to c_h \quad \text{as } m \to \infty.$$

By [12, Theorem 4.3] and Lemma 2.1 again, there exist a sequence $\{w_m\}$ in E_h and a sequence $\{\xi_m\}$ such that

$$I_h(w_m) \to c_h,\ I_h'(w_m) \to 0 \quad \text{and} \quad \|w_m - \xi_m u_m\|_{E_h} \to 0. \tag{37}$$

(The fact that [12, Theorem 4.3] applies in situations like ours was pointed out in [19].)

CLAIM 2.7.1. *$|\xi_m|$ has a positive lower bound.*

Otherwise, since $\{u_m\}$ is bounded in E_h, we have, after passing to a subsequence if necessary, $w_m \to 0$ in E_h. Hence $I_h(w_m) \to 0 = c_h$, which contradicts Lemma 2.6. This proves Claim 2.7.1.

By (37), for large m,

$$\begin{aligned} c_h + 1 + \|w_m\|_{E_h} &\ge I_h(w_m) - \frac{1}{q+1} I_h'(w_m) w_m \\ &\ge \left(\frac{1}{2} - \frac{1}{q+1}\right) \|w_m\|_{E_h}^2. \end{aligned}$$

It follows that $\{w_m\}$ is bounded in E_h. Hence by (37), $\{\xi_m\}$ is bounded. Furthermore, along a subsequence if necessary, $w_m \to$ some w_0 weakly in E_h, strongly in

$L^{\tau}_{\mathrm{loc}}(R^n)$, $1<\tau<2n/(n-2)^+$, and almost everywhere in R^n; by (37) again, w_0 is a weak and hence classical solution of equation (7).

If $w_0 \not\equiv 0$, then $w_0 \in M_h$. By Fatou's lemma,

$$\begin{aligned}
c_h &= \lim_{m\to\infty}\left(I_h(w_m) - \frac{1}{q+1}I_h'(w_m)w_m\right)\\
&= \lim_{m\to\infty}\left[\left(\frac{1}{q+1}-\frac{1}{p+1}\right)\int_{R^n}K(hx)|w_m|^{p+1}\,dx + \left(\frac{1}{2}-\frac{1}{q+1}\right)\|w_m\|^2_{E_h}\right]\\
&\geq \left(\frac{1}{q+1}-\frac{1}{p+1}\right)\int_{R^n}K(hx)|w_0|^{p+1}\,dx + \left(\frac{1}{2}-\frac{1}{q+1}\right)\|w_0\|^2_{E_h}\\
&= I_h(w_0) \geq c_h.
\end{aligned}$$

This implies that $I_h(w_0)=c_h$ and hence $|w_0|$ is a positive ground state of (7), and we are done.

Now we show that $w_0 \not\equiv 0$ for small h. Otherwise,

$$w_m \to 0 \quad \text{in } L^{\tau}_{\mathrm{loc}}(R^n), \quad 1<\tau<\frac{2n}{(n-2)^+}. \tag{38}$$

Claim 2.7.2. *$\int_{R^n}|u_m|^{p+1}\,dx$ has a positive lower bound.*

Otherwise, by passing to a subsequence if necessary and by Hölder's inequality, we have that the L^{p+1} and L^{q+1} norms of u_m converge to 0 as $m\to\infty$. Thus this is also true for w_m by (37) and the boundedness of ξ_m. Now we are led to

$$\begin{aligned}
c_h &= \lim_{m\to\infty}\left[I_h(w_m) - \frac{1}{2}I_h'(w_m)w_m\right]\\
&= \lim_{m\to\infty}\int_{R^n}\left[\left(\frac{1}{2}-\frac{1}{p+1}\right)K(hx)|w_m|^{p+1} + \left(\frac{1}{2}-\frac{1}{q+1}\right)Q(hx)|w_m|^{q+1}\right]dx\\
&\leq \lim_{m\to\infty}\left[\left(\frac{1}{2}-\frac{1}{p+1}\right)\|K\|_{L^\infty}\int_{R^n}|w_m|^{p+1}\,dx\right.\\
&\quad \left. + \left(\frac{1}{2}-\frac{1}{q+1}\right)\|Q\|_{L^\infty}\int_{R^n}|w_m|^{q+1}\,dx\right]\\
&= 0.
\end{aligned}$$

This contradicts Lemma 2.6 and proves Claim 2.7.2.

By assumption (36) and Lemma 2.4, we can choose $\epsilon>0$ so small that

$$c^{\epsilon} := c\,(V_\infty-\epsilon, K_\infty+\epsilon, Q_\infty+\epsilon) > \inf_{s\in R^n} C\,(s) \tag{39}$$

Take a constant ρ sufficiently large such that for $|x|>\rho$,

$$V(x) > V_\infty-\epsilon,\ K(x) < K_\infty+\epsilon,\ Q(x) < Q_\infty+\epsilon. \tag{40}$$

Let M^ϵ denote the solution manifold for the equation

$$\Delta v - (V_\infty-\epsilon)v + (K_\infty+\epsilon)|v|^{p-1}v + (Q_\infty+\epsilon)|v|^{q-1}v = 0, \quad x\in R^n. \tag{41}$$

From the proof of Lemma 2.1, there exists $\alpha_m>0$ such that $\alpha_m u_m \in M^\epsilon$. It is easy to see that for some $\kappa>0$ independent of m,

$$\alpha_m^{p+1}\int_{R^n}(K_\infty+\epsilon)|u_m|^{p+1}\,dx + \alpha_m^{q+1}\int_{R^n}(Q_\infty+\epsilon)|u_m|^{q+1}\,dx$$

$$
\begin{aligned}
&= \alpha_m^2 \int_{R^n} (|\nabla u_m|^2 + (V_\infty - \epsilon)|u_m|^2)\, dx \\
&\le \kappa \alpha_m^2 \|u_m\|_{E_h}^2 \;=\; \kappa \alpha_m^2 .
\end{aligned}
$$

This and Claim 2.7.2 imply the boundedness of α_m.

Now observe that by (40),

$$
\begin{aligned}
c_h &= \lim_{m\to\infty} \max_{t\ge 0} I_h(tu_m) \\
&\ge \limsup_{m\to\infty} I_h(\alpha_m u_m) \\
&= \limsup_{m\to\infty} \Bigg[\frac{1}{2}\int_{R^n} (|\nabla(\alpha_m u_m)|^2 + V(hx)(\alpha_m u_m)^2)\, dx \\
&\qquad - \frac{1}{p+1}\int_{R^n} K(hx)|\alpha_m u_m|^{p+1}\, dx - \frac{1}{q+1}\int_{R^n} Q(hx)|\alpha_m u_m|^{q+1}\, dx \Bigg] \\
&\ge \limsup_{m\to\infty} \Bigg[\frac{1}{2}\int_{R^n} (|\nabla(\alpha_m u_m)|^2 + (V_\infty - \epsilon)(\alpha_m v_m)^2)\, dx \\
&\qquad - \frac{1}{p+1}\int_{R^n} (K_\infty + \epsilon)|\alpha_m v_m|^{p+1}\, dx - \frac{1}{q+1}\int_{R^n} (Q_\infty + \epsilon)|\alpha_m v_m|^{q+1}\, dx \\
&\qquad + \frac{1}{2}\int_{B_{\rho/h}} (V(hx) - (V_\infty - \epsilon))(\alpha_m u_m)^2\, dx \\
&\qquad - \frac{1}{p+1}\int_{B_{\rho/h}} (K(hx) - (K_\infty + \epsilon))|\alpha_m u_m|^{p+1}\, dx \\
&\qquad - \frac{1}{q+1}\int_{B_{\rho/h}} (Q(hx) - (Q_\infty + \epsilon))|\alpha_m u_m|^{q+1}\, dx \Bigg], \qquad (42)
\end{aligned}
$$

where $B_{\rho/h}$ is the ball in R^n centered at 0 with radius ρ/h.

On the other hand, by the Sobolev embedding theorem, for $1 < \tau < 2n/(n-2)^+$,

$$
\|w_m - \xi_m u_m\|_{L^\tau} \le C(n,\, \tau)\|w_m - \xi_m u_m\|_{E_h} \to 0.
$$

Therefore, by Claim 2.7.1 and (38), we have that $\|u_m\|_{L^\tau(B_{\rho/h})} \to 0$ as $m \to \infty$. This fact, together with (42) and the boundedness of α_m, implies $c_h \;\ge\; c^\epsilon$, which is impossible for small h according to Lemma 2.6 and (39). □

The following result gives several sufficient conditions for existence which are more specific than condition (36).

COROLLARY 2.8. *If one of the following hypotheses holds true, then for small h, equation* (7) (*and hence* (4)) *has a positive ground-state solution.* (*Recall that V_∞, K_∞, and Q_∞ are defined in* (35).)

1. $V_\infty = \sup_{x\in R^n} V(x)$, $K_\infty = \inf_{x\in R^n} K(x)$, *and* $Q_\infty = \inf_{x\in R^n} Q(x)$.
2. *There exists a point $s_0 \in R^n$ such that*

$$
V_\infty \ge V(s_0), \qquad K_\infty \le K(s_0), \quad \textit{and} \quad Q_\infty \le Q(s_0),
$$

with one of the above inequalities being strict.

3. *There exists a point $s_0 \in R^n$ such that*

$$
\frac{V_\infty^{(2p+2+n-np)/(2p-2)}}{K_\infty^{2/(p-1)}} \ge \frac{V^{(2p+2+n-np)/(2p-2)}(s_0)}{K^{2/(p-1)}(s_0)},
$$

$$
\frac{Q_\infty}{V_\infty^{(p-q)/(p-1)} K_\infty^{(q-1)/(p-1)}} \le \frac{Q(s_0)}{V^{(p-q)/(p-1)}(s_0)\, K^{(q-1)/(p-1)}(s_0)},
$$

with one of the above inequalities being strict.

4. $K_\infty = 0$ *and* $Q_\infty < 0$. *In this case, h need not be small.*

Remark. If $V_\infty = +\infty$, then it is already proved in [19, Theorem 1.7] that (4) has a positive ground state for every $h > 0$. See also a recent preprint by Bartsch and (Z. Q.) Wang, "Existence and multiplicity results for some superlinear elliptic problems on R^n," which weakens this condition. However, our result here is not covered by [19] or any other papers to our knowledge.

Proof. By using Lemmas 2.2 and 2.5, it is easy to verify that each of the conditions 1–3 is sufficient to guarantee that either $c_\infty > \inf_{s\in R^n} C(s)$ or $V(x)$, $K(x)$, and $Q(x)$ are all constant functions. The desired existence follows either from Theorem 2.7 or, in the case where $V(x)$, $K(x)$, and $Q(x)$ are constant functions, from [20, Theorem 4.23].

Under condition 4, note that by Lemma 2.2, $c_{\alpha(s)} \geq c(1,1,0) > 0$ if $|s|$ is sufficiently large. That is, $c_{\alpha(s)}$ has a positive lower bound. Then (33) implies that $c_\infty = +\infty$. Replacing every V_∞ in the proof of Theorem 2.7 by a large number, we see that (4) has a positive ground state for every h. □

In the particular case of equation (1), i.e., $Q(x) \equiv 0$ in equation (4), Corollary 2.8 immediately implies the following.

Corollary 2.9. *If*

$$\frac{\liminf_{|x|\to\infty} V^{(2p+2+n-np)/(2p-2)}(x)}{\limsup_{|x|\to\infty} K^{2/(p-1)}(x)} > \inf_{x\in R^n} \frac{V^{(2p+2+n-np)/(2p-2)}(x)}{K^{2/(p-1)}(x)}, \tag{43}$$

then for small h, equation (1) *has a positive ground-state solution.*

3. Concentration of ground states. In this and the following sections, v_h is always referred to a positive ground state of (7), and $u_h(x) := v_h(x/h)$ is always a positive ground state of (4). We shall always assume that they are related in this way.

Theorem 3.1. *Under condition* (36), *for every sequence* $\{h'_k\} \to 0^+$, *there exists a subsequence* $\{h_k\}$ *such that a sequence of positive ground states* $\{u_{h_k}(x)\}$ *of* (4) *concentrates at a global minimum point* x_0 *of* $C(s)$ *in the following sense: for each small positive* h_k, $u_{h_k}(x)$ *has a unique maximum point* x_k *with* $\lim_{h_k\to 0^+} x_k = x_0$; *moreover, for each positive* δ *and large* k,

$$\max_{|x-x_0|\leq\delta} u_{h_k}(x) > C_1 \tag{44}$$

and

$$u_{h_k}(x) \leq C_2 \left|\frac{x-x_k}{h_k}\right|^{\frac{1-n}{2}} \exp\left(-\overline{V}^{\frac{1}{2}}\left|\frac{x-x_k}{h_k}\right|\right) \quad \text{for } x \in R^n, \tag{45}$$

where C_1 *and* C_2 *depend only on* n, p, q, $\overline{V} := \inf V$, $\widehat{K} := \sup K$, *and* $\widehat{Q} := \sup Q$.

As an immediate consequence of this and Lemma 2.5, we have the following result.

Corollary 3.2. *Assume* (43). *For every sequence* $\{h'_k\} \to 0^+$, *there exists a subsequence* $\{h_k\}$ *such that a sequence of positive ground states* $\{u_{h_k}(x)\}$ *of* (1) *concentrates at a global minimum point* x_0 *of*

$$g(x) := \frac{V^{(2p+2+n-np)/(2p-2)}(x)}{K^{2/(p-1)}(x)}$$

in the sense specified in the statement of Theorem 3.1.

The proof of this theorem will be lengthy but will be along the main lines of the proof of the corresponding result in [22]. We shall first show that there exists a sequence of points $\{y_{h_k}\}$ in R^n such that (i) most of the "mass" of v_{h_k} is contained in a ball (of fixed size) centered at y_{h_k} and (ii) $h_k y_{h_k}$ is bounded. This will be done in Lemmas 3.3 and 3.4. Then in Lemma 3.5, we show that (i) any limit point of $h_k y_{h_k}$ is a global minimum point of the ground-energy function $C(s)$ and (ii) $w_k(x) := v_{h_k}(x + y_{h_k}) = u_{h_k}(h_k x + h_k y_{h_k})$ converges in $H^1(R^n)$. After finishing these steps, the theorem will follow from modifying the arguments in the proof of [22, Theorem 2.1] (which is from (2.15) to the end in that paper). We shall not give the details of the easy modifications.

Now we proceed to prepare the first of these lemmas. Observe that for any v on the solution manifold M_h,

$$\begin{aligned} I_h(v) = {} & \left(\frac{1}{2} - \frac{1}{q+1}\right) \int_{R^n} (|\nabla v|^2 + V(hx)v^2)\, dx \\ & + \left(\frac{1}{q+1} - \frac{1}{p+1}\right) \int_{R^n} K(hx)|v|^{p+1}\, dx. \end{aligned} \tag{46}$$

Define a *measure* μ_h by

$$\mu_h(\Omega) = \int_\Omega \left[\left(\frac{1}{2} - \frac{1}{q+1}\right)(|\nabla v_h|^2 + V(hx)v_h^2) + \left(\frac{1}{q+1} - \frac{1}{p+1}\right) K(hx)|v_h|^{p+1}\right] dx.$$

By Lemma 2.6, along a subsequence if necessary, as $h \to 0$,

$$\mu_h(R^n) = c_h \to \tilde{c} \leq \inf_{s \in R^n} C\,(s), \tag{47}$$

where $\tilde{c} \geq \bar{c} > 0$. By the concentration-compactness lemma of P. L. Lions in [11, part 1] (or see [21]), there are three possibilities:

1 (*compactness*). There exists a sequence $\{y_{h_k}\}$ that satisfies the following: for any $\epsilon > 0$, there is a $\rho > 0$ such that

$$\int_{B_\rho(y_{h_k})} d\mu_{h_k} \geq \tilde{c} - \epsilon. \tag{48}$$

2 (*vanishing*). There exists a sequence $\{h_k\}$ that tends to zero such that for all $\rho > 0$,

$$\lim_{h_k \to 0^+} \sup_{y \in R^n} \int_{B_\rho(y)} d\mu_{h_k} = 0;$$

3 (*dichotomy*). There exist a constant $\tilde{c}'$ with $0 < \tilde{c}' < \tilde{c}$, sequences $\{\rho_{h_k}\} \to \infty$ and $\{y_{h_k}\} \subset R^n$, and two nonnegative measures $\mu^1_{h_k}$ and $\mu^2_{h_k}$ such that

$$0 \leq \mu^1_{h_k} + \mu^2_{h_k} \leq \mu_{h_k},$$

$$\operatorname{supp}(\mu^1_{h_k}) \subset B_{\rho_{h_k}}(y_{h_k}), \qquad \operatorname{supp}(\mu^2_{h_k}) \subset B^c_{2\rho_{h_k}}(y_{h_k}),$$

$$\mu^1_{h_k}(R^n) \to \tilde{c}', \qquad \mu^2_{h_k}(R^n) \to \tilde{c} - \tilde{c}'.$$

LEMMA 3.3. *Neither vanishing* (2) *nor dichotomy* (3) *occurs.*

Proof.

CLAIM 3.3.1. *Vanishing* (2) *does not occur.*

Otherwise, $v_{h_k} \to 0$ in L^τ for each τ with $2 < \tau < 2n/(n-2)^+$ (see [11, part 2] or [4]). Then

$$\begin{aligned} 0 &= \lim_{k\to\infty} \left(\frac{1}{2} - \frac{1}{p+1}\right) \widehat{K} \int_{R^n} v_{h_k}^{p+1}\,dx + \left(\frac{1}{2} - \frac{1}{q+1}\right) \widehat{Q} \int_{R^n} v_{h_k}^{q+1}\,dx \\ &\geq \limsup_{k\to\infty} \left(\frac{1}{2} - \frac{1}{p+1}\right) \int_{R^n} K(h_k x) v_{h_k}^{p+1}\,dx + \left(\frac{1}{2} - \frac{1}{q+1}\right) \int_{R^n} Q(h_k x) v_{h_k}^{q+1}\,dx \\ &= \limsup_{k\to\infty} c_{h_k} \ \geq\ \bar{c}\ >\ 0. \end{aligned}$$

This contradiction proves Claim 3.3.1.

CLAIM 3.3.2. *Dichotomy* (3) *does not occur.*

Otherwise, take $\phi_h \in C_0^1(R^n)$ such that $\phi_h \equiv 1$ in $B_{\rho_h}(y_h)$, $\phi_h \equiv 0$ in $B^c_{2\rho_h}(y_h)$, and $0 \leq \phi_h \leq 1$, $|\nabla \phi_h| \leq 2/\rho_h$.

Write

$$v_h = \phi_h v_h + (1-\phi_h) v_h =: v_{1h} + v_{2h},$$

where v_{1h} and v_{2h} are defined in the last equality. Then as $h_k \to 0$,

$$\begin{aligned} I_{h_k}(v_{1h_k}) &\geq \mu_{h_k}(B_{\rho_{h_k}}(y_{h_k}))\ \geq\ \mu^1_{h_k}(B_{\rho_h}(y_{h_k})) \\ &= \mu^1_{h_k}(R^n)\ \to\ \tilde{c}' \end{aligned} \tag{49}$$

and

$$\begin{aligned} I_{h_k}(v_{2h_k}) &\geq \mu_{h_k}(B^c_{2\rho_{h_k}}(y_{h_k}))\ \geq\ \mu^2_{h_k}(B^c_{2\rho_{h_k}}(y_{h_k})) \\ &= \mu^2_{h_k}(R^n)\ \to\ \tilde{c} - \tilde{c}'. \end{aligned} \tag{50}$$

Let $\Omega_h = B_{2\rho_h}(y_h) \backslash B_{\rho_h}(y_h)$. Then

$$\begin{aligned} \left(\frac{1}{2} - \frac{1}{q+1}\right) \int_{\Omega_{h_k}} (|\nabla\ v_{h_k}|^2 + V(h_k x) v_{h_k}^2)\,dx + \left(\frac{1}{q+1} - \frac{1}{p+1}\right) \int_{\Omega_{h_k}} K(h_k x) v_{h_k}^{p+1}\,dx \\ = \mu_{h_k}(\Omega_{h_k}) \\ = \mu_{h_k}(R^n) - \mu_{h_k}(B_{\rho_{h_k}}(y_{h_k})) - \mu_{h_k}(B^c_{2\rho_{h_k}}(y_{h_k})) \\ \leq \mu_{h_k}(R^n) - \mu^1_{h_k}(R^n) - \mu^2_{h_k}(R^n) \\ \to 0. \end{aligned} \tag{51}$$

Thus by the Sobolev embedding theorem, we have

$$\int_{\Omega_{h_k}} (v_{h_k}^{p+1} + v_{h_k}^{q+1})\,dx \to 0 \quad \text{as } h_k \to 0^+.$$

Consequently,

$$\int_{R^n} K(h_k x) v_{h_k}^{p+1}\,dx = \int_{R^n} K(h_k x)(v_{1h_k} + v_{2h_k})^{p+1}\,dx$$

$$\begin{aligned}
&= \int_{B_{\rho_{h_k}}(y_{h_k})} K(h_k x) v_{1h_k}^{p+1}\, dx + \int_{\Omega_{h_k}} K(h_k x) v_{h_k}^{p+1}\, dx \\
&\quad + \int_{B^c_{2\rho_{h_k}}(y_{h_k})} K(h_k x) v_{2h_k}^{p+1}\, dx \\
&= \int_{R^n} K(h_k x) v_{1h_k}^{p+1}\, dx + \int_{R^n} K(h_k x) v_{2h_k}^{p+1}\, dx + o(1). \qquad (52)
\end{aligned}$$

Similarly,

$$\int_{R^n} Q(h_k x) v_{h_k}^{q+1}\, dx = \int_{R^n} Q(h_k x) v_{1h_k}^{q+1}\, dx + \int_{R^n} Q(h_k x) v_{2h_k}^{q+1}\, dx + o(1). \qquad (53)$$

Next, observe that

$$\begin{aligned}
\int_{R^n} (|\nabla v_{h_k}|^2 + V(h_k x) v_{h_k}^2)\, dx &= \int_{R^n} (|\nabla v_{1h_k}|^2 + V(h_k x) v_{1h_k}^2)\, dx \\
&\quad + \int_{R^n} (|\nabla v_{2h_k}|^2 + V(h_k x) v_{2h_k}^2)\, dx + J_{h_k}, \qquad (54)
\end{aligned}$$

where $J_{h_k} := 2\int_{R^n} (\nabla v_{1h_k} \cdot \nabla v_{2h_k} + V(h_k x)\, v_{1h_k} v_{2h_k})\, dx \to 0$ as $h_k \to 0$ because of (51).

Now (49)–(50) and (52)–(54) imply that

$$\begin{aligned}
\tilde{c} &= \lim_{h_k \to 0^+} I_{h_k}(v_{h_k}) \\
&= \lim_{h_k \to 0^+} (I_{h_k}(v_{1h_k}) + I_{h_k}(v_{2h_k}) + o(1)) \\
&\geq \liminf_{h_k \to 0^+} I_{h_k}(v_{1h_k}) + \liminf_{h_k \to 0^+} I_{h_k}(v_{2h_k}) \\
&\geq \tilde{c}' + (\tilde{c} - \tilde{c}') = \tilde{c}.
\end{aligned}$$

Therefore,

$$\lim_{h_k \to 0^+} I_{h_k}(v_{1h_k}) = \tilde{c}', \qquad \lim_{h_k \to 0^+} I_{h_k}(v_{2h_k}) = \tilde{c} - \tilde{c}'. \qquad (55)$$

Let

$$J^1_{h_k} := \int_{R^n} (|\nabla v_{1h_k}|^2 + V(h_k x) v_{1h_k}^2)\, dx - \int_{R^n} K(h_k x) v_{1h_k}^{p+1}\, dx - \int_{R^n} Q(h_k x) v_{1h_k}^{q+1}\, dx$$

and

$$J^2_{h_k} := \int_{R^n} (|\nabla v_{2h_k}|^2 + V(h_k x) v_{2h_k}^2)\, dx - \int_{R^n} K(h_k x) v_{2h_k}^{p+1}\, dx - \int_{R^n} Q(h_k x) |v_{2h_k}|^{q+1}\, dx.$$

By the fact that $v_h \in M_h$ and by equalities (52)–(54), we get

$$J^1_{h_k} = -J^2_{h_k} + o(1). \qquad (56)$$

Now we conclude our proof of Claim 3.3.2 by showing that (56) is not true. We discuss all three possible cases and show that each leads to a contradiction.

For simplicity of notation, let

$$A_1 := \textstyle\int_{R^n} K(h_k x) v_{1h_k}^{p+1}\, dx \quad \text{and} \quad B_1 := \int_{R^n} Q(h_k x) v_{1h_k}^{q+1}\, dx.$$

Take $\theta_1 > 0$ such that $\theta_1 v_{1h} \in M_h$. That is,

$$\theta_1^{p+1} A_1 + \theta_1^{q+1} B_1 = \theta_1^2 \int_{R^n} (|\nabla v_{1h}|^2 + V(hx) v_{1h}^2)\, dx.$$

Case 1. After passing to a subsequence, $J_{h_k}^1 \leq 0$. In this case, we have

$$\theta_1^{p-1} A_1 + \theta_1^{q-1} B_1 = \int_{R^n} (|\nabla v_{1h_k}|^2 + V(h_k x) v_{1h_k}^2)\, dx \leq A_1 + B_1.$$

Thus $\theta_1 \leq 1$ (see (15)) and hence by (55), as $h_k \to 0^+$,

$$c_{h_k} \leq I_{h_k}(\theta_1 v_{1h_k}) \leq I_{h_k}(v_{1h_k}) \to \tilde{c}' < \tilde{c}.$$

This is absurd because $c_{h_k} \to \tilde{c} > \tilde{c}'$.

Case 2. After passing to a subsequence, $J_{h_k}^2 \leq 0$. In this case, we will be led to a contradiction again as in Case 1.

Case 3. After passing to a subsequence, $J_{h_k}^1 > 0$ and $J_{h_k}^2 > 0$. From (56), it follows that $J_{h_k}^1 = o(1)$ and $J_{h_k}^2 = o(1)$. If $\theta_1 \leq 1 + o(1)$, we are done by arguments similar to those in the proof for Case 1. Now suppose that $\lim_{h_k \to 0^+} \theta_1 = \theta_0 > 1$. We claim that along a subsequence, $\lim_{h_k \to 0^+} (A_1 + B_1) > 0$. Otherwise,

$$\lim_{h_k \to 0^+} \int_{R^n} (|\nabla v_{1h_k}|^2 + V(h_k x) v_{1h_k}^2)\, dx \leq \lim_{h_k \to 0^+} J_{h_k}^1 = 0,$$

which implies that $\tilde{c}' = \lim_{h_k \to 0^+} I_{h_k}(v_{1h_k}) = 0$. This is absurd.

Now observe that

$$\begin{aligned} 0 = \lim_{h_k \to 0^+} J_{h_k}^1 &= \lim_{h_k \to 0^+} (\theta_1^{p-1} A_1 + \theta_1^{q-1} B_1 - A_1 - B_1) \\ &\geq \lim_{h_k \to 0^+} (\theta_1^{q-1} - 1)(A_1 + B_1) = (\theta_0^{q-1} - 1) \lim_{h_k \to 0^+} (A_1 + B_1) \\ &> 0. \end{aligned}$$

We are led to a contradiction again. This proves Claim 3.3.2 and Lemma 3.3. □

Henceforth in this section, the sequence $\{y_{h_k}\}$ is always referred to the one obtained in (48).

Let $w_k(x) := v_{h_k}(x + y_{h_k}) = u_{h_k}(h_k x + h_k y_{h_k})$. Then $w_k(x)$ is a positive ground state of

$$\Delta w_k - V(h_k x + h_k y_{h_k}) w_k + K(h_k x + h_k y_{h_k}) w_k^p + Q(h_k x + h_k y_{h_k}) w_k^q = 0. \tag{57}$$

LEMMA 3.4. *If* (36) *holds, then the sequence* $\{h_k y_{h_k}\}$ *is bounded as* h_k *tends to zero.*

Proof. Suppose that after passing to a subsequence, $h_k y_{h_k} \to \infty$. Since c_{h_k} is bounded, so is $w_k := w_{h_k}$ in H^1. Therefore, along a subsequence, $w_k \to$ some w_0 weakly in H^1, strongly in L^τ_{loc}, where $1 < \tau < 2n/(n-2)^+$, and almost everywhere in R^n. By the compactness condition (48), for any $\epsilon > 0$, there exists $\rho > 0$ such that

$$\left(\frac{1}{2} - \frac{1}{q+1}\right) \int_{B_\rho^c} (|\nabla w_k|^2 + \overline{V}\, w_k^2)\, dx \leq \mu_{h_k}(B_\rho^c(y_{h_k})) < \epsilon.$$

By this and the Sobolev embedding theorem, we have that

$$w_k \to w_0 \quad \text{strongly in } L^\tau, \quad 1 < \tau < \frac{2n}{(n-2)^+}. \tag{58}$$

Observe that

$$\begin{aligned}\left(\frac{1}{2}-\frac{1}{p+1}\right)&\int_{R^n}\widehat{K}w_0^{p+1}\,dx+\left(\frac{1}{2}-\frac{1}{q+1}\right)\int_{R^n}\widehat{Q}w_0^{q+1}\,dx\\&\geq \limsup_{h_k\to 0^+}\left[\left(\frac{1}{2}-\frac{1}{p+1}\right)\int_{R^n}K(h_kx+h_ky_{h_k})w_k^{p+1}\,dx\right.\\&\qquad\left.+\left(\frac{1}{2}-\frac{1}{q+1}\right)\int_{R^n}Q(h_kx+h_ky_{h_k})w_k^{q+1}\,dx\right]\\&=\limsup_{h_k\to 0^+} c_{h_k}\\&\geq \bar{c}>0.\end{aligned}$$

Thus $w_0(x)$ is a nonzero function. Choose $\epsilon>0$ such that (39) holds. By (58) and the assumption $h_ky_{h_k}\to\infty$, we have

$$\Delta w_0-\left(V_\infty-\frac{\epsilon}{2}\right)w_0+\left(K_\infty+\frac{\epsilon}{2}\right)w_0^p+\left(Q_\infty+\frac{\epsilon}{2}\right)w_0^q\geq 0\quad\text{in } H^{-1}.$$

In particular,

$$\int_{R^n}(|\nabla w_0|^2+(V_\infty-\epsilon)w_0^2)\,dx<\int_{R^n}(K_\infty+\epsilon)|w_0|^{p+1}\,dx+\int_{R^n}(Q_\infty+\epsilon)|w_0|^{p+1}\,dx$$

since $w_0\not\equiv 0$. Take $\theta>0$ such that $\theta w_0\in M^\epsilon$, the solution manifold for equation (41). Then $\theta<1$ by the above inequality. Let

$$A:=\textstyle\int_{R^n}K(h_kx+h_ky_{h_k})w_{h_k}^{p+1}\,dx\quad\text{and}\quad B:=\int_{R^n}Q(h_kx+h_ky_{h_k})w_{h_k}^{q+1}\,dx.$$

By (58) and the assumption that $h_ky_{h_k}\to\infty$, we have

$$\limsup_{h_k\to 0^+}A\leq\int_{R^n}(K_\infty+\epsilon)w_0^{p+1}\,dx\quad\text{and}\quad\limsup_{h_k\to 0^+}B\leq\int_{R^n}(Q_\infty+\epsilon)w_0^{q+1}\,dx.$$

These and (48) imply that

$$\begin{aligned}c^\epsilon&:=c\,(V_\infty-\epsilon,K_\infty+\epsilon,Q_\infty+\epsilon)\\&\leq\frac{1}{2}\theta^2\int_{R^n}(|\nabla w_0|^2+(V_\infty-\epsilon)w_0^2)\,dx\\&\quad-\frac{\theta^{p+1}}{p+1}\int_{R^n}(K_\infty+\epsilon)w_0^{p+1}\,dx-\frac{\theta^{q+1}}{q+1}\int_{R^n}(Q_\infty+\epsilon)w_0^{q+1}\,dx\\&\leq\liminf_{h_k\to 0+}\left[\frac{1}{2}\theta^2\int_{R^n}(|\nabla w_k|^2+V(h_kx+h_ky_m)w_k^2)\,dx\right.\\&\qquad-\frac{\theta^{p+1}}{p+1}\int_{R^n}K(h_kx+h_ky_m)w_k^{p+1}\,dx\\&\qquad\left.-\frac{\theta^{q+1}}{q+1}\int_{R^n}Q(h_kx+h_ky_m)w_k^{q+1}\,dx\right]\\&=\liminf_{h_k\to 0+}f(\theta),\end{aligned}$$

where $f(\theta):=(1/2)\theta^2(A+B)-(1/(p+1))\theta^{p+1}A-(1/(q+1))\theta^{q+1}B$. Noting that $A+B>0$, we can easily show that $(d/d\theta)f(\theta)=(A+B)\theta-A\theta^p-B\theta^q>0$ for $\theta\in(0,1)$. Consequently, $f(\theta)<f(1)$ for $\theta\in(0,1)$. This and Lemma 2.6 yield

$$c^\epsilon\leq\liminf_{h_k\to 0+}f(1)=\liminf_{h_k\to 0+}c_{h_k}\leq\inf_{s\in R^n}C\,(s),$$

a contradiction to (39). ∎

Now from Lemma 3.4 and its proof, we know that for any sequence $\{h'_k\} \to 0$, there exists a subsequence $\{h_k\}$ such that $\bar{x}_k := h_k y_{h_k} \to x_0$, and $w_k \to w_0$ weakly in H^1, where $w_0 \geq 0, \not\equiv 0$; moreover, (58) holds.

LEMMA 3.5. $C(x_0) = \inf_{s\in R^n} C(s)$. *Furthermore, $w_k \to w_0$ strongly in H^1.*

Proof. By (58) and the elliptic regularity theory, as $k \to \infty$, $w_k \to w_0$ in C^2_{loc} and

$$\Delta w_0 - V(x_0)w_0 + K(x_0)w_0^p + Q(x_0)w_0^q = 0, \quad x \in R^n.$$

Consequently, by (48) and (58) again,

$$\begin{aligned}
&\inf_{s\in R^n} C(s) \\
&\leq C(x_0) \\
&\leq \left(\frac{1}{2} - \frac{1}{q+1}\right)\int_{R^n} (|\nabla w_0|^2 + V(x_0)w_0^2)\,dx + \left(\frac{1}{q+1} - \frac{1}{p+1}\right)\int_{R^n} K(x_0)w_0^{p+1}\,dx \\
&\leq \liminf_{k\to\infty} \left[\left(\frac{1}{2} - \frac{1}{q+1}\right)\int_{R^n} (|\nabla w_k|^2 + V(h_k x + \bar{x}_k)w_k^2)\,dx \right.\\
&\quad \left. + \left(\frac{1}{q+1} - \frac{1}{p+1}\right)\int_{R^n} K(h_k x + \bar{x}_k)w_k^{p+1}\,dx\right] \\
&= \liminf_{k\to\infty} c_{h_k} \leq \inf_{s\in R^n} C(s).
\end{aligned}$$

This implies that $C(x_0) = \inf_{s\in R^n} C(s)$. From the above inequalities and (58), we see that

$$\lim_{k\to\infty}\int_{R^n} (|\nabla w_k|^2 + V(h_k x + \bar{x}_k)w_k^2)\,dx = \int_{R^n} (|\nabla w_0|^2 + V(x_0)w_0^2)\,dx.$$

From this and by arguments involving the application of Fatou's lemma on the complement of large balls, it is easy to show that $w_k \to w_0$ in H^1. ∎

As mentioned earlier, Theorem 3.1 follows from this lemma and the arguments in [22]. We omit the details.

Remarks. 1. Condition (36) is used in the proof of Theorem 3.1 to ensure the existence of ground states and to show the boundedness of $\{h_k y_{h_k}\}$ in Lemma 3.4. If we do not have condition (36) but know that a sequence $\{u_{h_k}\}$ of positive ground states of (4) exists with each having a local maximum point moving toward a point x_0 as $h_k \to 0$, then by modifying our proof above, it is easy to show that x_0 is a global minimum point of the ground-energy function $C(s)$ and that $\{u_{h_k}\}$ concentrates at x_0 in the sense specified in the statement of Theorem 3.1. See also [22, Theorem 2.3].

2. From this we see that in the even worse scenario where $C(s)$ does not have a minimum point, the positive ground states, if any, do not concentrate as h shrinks (they move off to infinity).

3. As pointed out in section 1, unlike (1), in the general case (4), it is impossible to have an explicit formula for $C(s)$, and hence we cannot explicitly express the location of concentration of ground states in terms of the potential functions $V(x)$, $K(x)$, and $Q(x)$. We believe that if the exponent p in (4) is close to the critical exponent $(n+2)/(n-2)^+$, then the concentration points of ground states are close to maximum points of $K(x)$, and if the other exponent q is close to 1, then these concentration points are close to minimum points of $(V(x) - Q(x))^{(2p+2+n-np)/(2p-2)}/K^{2/(p-1)}(x)$. The techniques developed in [23] may be useful in proving this.

4. Necessary condition for location of concentration. In this section, we assume that there are positive constants γ and C such that

$$|\nabla V(x)|, |\nabla K(x)|, |\nabla Q(x)| \leq C \exp(\gamma |x|). \tag{59}$$

THEOREM 4.1. *Suppose that a sequence of positive bound states $u_{h_k}(x)$ of* (4) *concentrates at a point x_0 in the following sense: for any $\epsilon > 0$, there exist positive constants ρ and N such that*

$$u_{h_k}(x) \leq \epsilon \quad \textit{for } k \geq N \quad \textit{and} \quad |x - x_0| \geq h_k \rho. \tag{60}$$

Then in any of the cases (a) $n = 1$, (b) $Q(x_0) \leq 0$ *in* R^n, *or* (c) $1 < q < p \leq n/(n-2)$ *and* $n > 2$, *the point x_0 is a critical point of the ground energy function $C(s)$: $\nabla C(x_0) = 0$. In particular, in the case of* (1) (*i.e.*, $Q \equiv 0$), $\nabla g(x_0) = 0$, *where* $g(x) = V^{(2p+2+n-np)/(2p-2)}(x)/K^{2/(p-1)}(x)$.

Remark. Each of (a)–(c) is used only to guarantee the uniqueness of positive decaying solutions of (63) below. Thus any condition that guarantees the uniqueness can be a substitute for either of (a)–(c).

Proof. We shall follow the main lines of the proof of the corresponding result in [22]. First, we show that $\{u_{h_k}\}$ is bounded in $L^\infty(R^n)$. This can be achieved by slightly modifying the argument in [22].

Next, let $w_k(x) = u_{h_k}(x_0 + h_k x)$. Then

$$\Delta w_k - V(x_0 + h_k x) w_k + K(x_0 + h_k x) w_k^p + Q(x_0 + h_k x) w_k^q = 0. \tag{61}$$

By (60), $w_k(x)$ decays to zero uniformly with respect to k. Then a simple comparison argument shows

$$w_k(x) \leq C_1 \exp(-C_2 |x|) \quad \text{for } x \in R^n, \tag{62}$$

where C_1 and C_2 are positive constants independent of k. By elliptic regularity theory, together with the boundedness of $\|w_k\|_{L^\infty}$, there exists a $w_0(x)$ in $C^2(R^n)$ such that, along a subsequence, $w_k(x)$ converges to $w_0(x)$ strongly in C^2_{loc}, satisfying (62) and

$$\Delta w_0(x) - V(x_0) w_0(x) + K(x_0) w_0^p(x) + Q(x_0) w_0^q(x) = 0. \tag{63}$$

By the maximum principle, $\max_{x \in R^n} w_k(x) \geq C_0 > 0$, where the constant C_0 is independent of k. This together with (63) implies that all the maximum points of $w_k(x)$ have to remain in a bounded domain for all k. Consequently, $\max_{x \in R^n} w_0(x) \geq C_0 > 0$, and by the strong-maximum principle, $w_0(x) > 0$. Furthermore, as mentioned in the proof of Lemma 2.3, in each of the cases (a)–(c), (63) has only one positive decaying solution (up to translation). Thus w_0 is a ground state of (63).

Multiplying (61) on both sides by $\nabla w_k(x)$ and integrating on $B_R = B_R(0)$, we have

$$\int_{B_R} \left[\Delta w_k \nabla w_k - \frac{1}{2} \nabla (V(x_0 + h_k x) w_k^2) + \frac{1}{2} h_k \nabla V(x_0 + h_k x) w_k^2 \right.$$
$$\left. + K(x_0 + h_k x) \frac{\nabla w_k^{p+1}}{p+1} + Q(x_0 + h_k x) \frac{\nabla w_k^{q+1}}{q+1} \right] dx = 0.$$

Therefore, by the divergence theorem, we have

$$h_k \int_{B_R} \left[\frac{1}{2} \nabla V(x_0 + h_k x) w_k^2 - \frac{1}{p+1} \nabla K(x_0 + h_k x) w_k^{p+1} \right.$$

$$
\begin{aligned}
&\qquad - \frac{1}{q+1}\nabla Q(x_0+h_k x) w_k^{q+1}\Big] \, dx \\
&= -\int_{B_R} \Delta w_k \nabla w_k \, dx + \int_{\partial B_R} \frac{1}{2}\Big[V(x_0+h_k x) w_k^2 \, \nu \\
&\qquad - \frac{1}{p+1} K(x_0+h_k x) w_k^{p+1} \nu - \frac{1}{q+1} Q(x_0+h_k x) w_k^{q+1} \nu \Big] \, dS \\
&= \int_{\partial B_R} \Big[-\nabla w_k \frac{\partial w_k}{\partial \nu} + \frac{|\nabla w_k|^2}{2}\nu + \frac{1}{2} V(x_0+h_k x) w_k^2 \, \nu \\
&\qquad - \frac{1}{p+1} K(x_0+h_k x) w_k^{p+1} \nu - \frac{1}{q+1} Q(x_0+h_k x) w_k^{q+1} \nu \Big] \, dS \\
&=: J_R,
\end{aligned}
$$

where ν stands for the unit outward normal to ∂B_R. Note that for each k,

$$
\begin{aligned}
\int_0^\infty |J_R| \, dR &\le \int_0^\infty dR \int_{\partial B_R} \Big[\frac{3}{2} |\nabla w_k|^2 + \frac{1}{2} V(x_0+h_k x) w_k^2 \\
&\qquad + \frac{1}{p+1} K(x_0+h_k x) w_k^{p+1} + \frac{1}{q+1} Q(x_0+h_k x) w_k^{q+1} \Big] \, dS \\
&\le \frac{3}{2} \int_{R^n} [|\nabla w_k|^2 + V(x_0+h_k x) w_k^2 + K(x_0+h_k x) w_k^{p+1} \\
&\qquad + Q(x_0+h_k x) w_k^{q+1}] \, dx,
\end{aligned}
$$

which is finite because u_{h_k} is a bound state. Consequently, there exists $R_m \to \infty$ such that $J_{R_m} \to 0$. Recall that we assume the growth condition (59) and that we have obtained estimate (62) for $w_k(x)$. By the Lebesgue dominated-convergence theorem,

$$
\begin{aligned}
&\frac{1}{2} \int_{R^n} \nabla V(x_0+h_k x) w_k^2 \, dx \\
&\qquad = \frac{1}{p+1} \int_{R^n} \nabla K(x_0+h_k x) w_k^{p+1} \, dx + \frac{1}{q+1} \int_{R^n} \nabla Q(x_0+h_k x) w_k^{q+1} \, dx.
\end{aligned}
$$

Letting $h_k \to 0$, we obtain

$$
\begin{aligned}
&\frac{1}{2} \nabla V(x_0) \int_{R^n} w_0^2(x) \, dx \\
&\qquad = \frac{1}{p+1} \nabla K(x_0) \int_{R^n} w_0^{p+1}(x) \, dx + \frac{1}{q+1} \nabla Q(x_0) \int_{R^n} w_0^{q+1}(x) \, dx. \qquad (64)
\end{aligned}
$$

Now the desired conclusion follows from Lemma 2.3 and the remark below it since w_0 is a ground state of (63). □

Note added in proof. Some of the papers mentioned in section 1 that concern the existence of concentrating bound states of (6) have already been published:

M. Del Pino and P. Felmer, "Local mountain passes for semilinear elliptic problems in unbounded domains," *Cal. Var. Partial Differential Equations*, 4 (1996), pp. 121–137.

C. Gui, "Existence of multi-bump solutions for nonlinear Schrödinger equations via variational method," *Comm. Partial Differential Equations*, 21 (1996), pp. 787–820.

Acknowledgment. We are very grateful to Professor Steve Rosencrans for his helpful suggestions and comments.

REFERENCES

[1] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations* I and II, Arch. Rational Mech. Anal., 82 (1983), pp. 313–345 and pp. 347–375.

[2] C. C. Chen and C. S. Lin, *Uniqueness of the ground state solutions of $\Delta u + f(u) = 0$ in R^n*, $n \geq 3$, Comm. Partial Differential Equations, 16 (1991), pp. 1549–1572.

[3] S. Coleman, V. Glaser, and A. Martin, *Action minima among solutions to a class of Euclidean scalar field equations*, Comm. Math. Phys., 58 (1978), pp. 211–221.

[4] V. Coti Zelati and P. H. Rabinowitz, *Homoclinic type solutions for a semilinear elliptic PDE on R^n*, Comm. Pure Appl. Math. 45 (1992), pp. 1217–1269.

[5] W.-Y. Ding and W.-M. Ni, *On the existence of positive entire solutions of a semilinear elliptic equation*, Arch. Rational Mech. Anal., 91 (1986), pp. 283–308.

[6] 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.

[7] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in R^n*, Adv. Math. Supplementary Stud., 7A (1981), pp. 369–402.

[8] 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.

[9] M. K. Kwong and L. Zhang, *Uniqueness of the positive solution of $\Delta u + f(u) = 0$ in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.

[10] Y. Li, *Remarks on a semilinear elliptic equation on R^n*, J. Differential Equations, 74 (1988), pp. 34–49.

[11] P. L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case parts* 1 and 2, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145 and pp. 223–283.

[12] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, Berlin, New York, 1989.

[13] Z. Nehari, *On a nonlinear differential equation arising in nuclear physics*, Proc. Roy. Irish Acad. Sect. A, 62 (1963), pp. 117–135.

[14] W.-M. Ni, *Recent progress in semilinear elliptic equations*, in RIMS Kokyuroku 678, T. Suzuki ed., Kyoto University Press, Kyoto, Japan, 1989, pp. 1–39.

[15] 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.

[16] Y. G. Oh, *Existence of semiclassical bound states of nonlinear Schrödinger equations with potentials of the class $(V)_a$*, Comm. Partial Differential Equations, 13 (1988), pp. 1499–1519.

[17] Y. G. Oh, *Correction to "Existence of semiclassical bound states of nonlinear Schrödinger equations with potentials of the class $(V)_a$,"* Comm. Partial Differential Equations, 14 (1989), pp. 833–834.

[18] Y. G. Oh, *On positive multi-lump bound states of nonlinear Schrödinger equations under multiple well potential*, Comm. Math. Phys., 131 (1990), pp. 223–253.

[19] P. H. Rabinowitz, *On a class of nonlinear Schrödinger equations*, Z. Angew. Math. Phys., 43 (1992), pp. 270–291.

[20] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[21] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, New York, 1990.

[22] X.-F. Wang, *On concentration of positive bound states of nonlinear Schrödinger equations*, Comm. Math. Phys., 153 (1993), pp. 229–244.

[23] X.-F. Wang, *On location of blow-up of ground states of semilinear elliptic equations in R^n involving critical Sobolev exponents*, J. Differential Equations, 127 (1996), pp. 148–173.

[24] B. Zeng, *On the existence and concentration of nonlinear Schrödinger equations*, Ph.D. thesis, Department of Mathematics, Tulane University, New Orleans, LA, 1995.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 656–668, May 1997

ERROR BOUNDS IN NONSMOOTH IMAGE DEBLURRING[*]

ALFRED S. CARASSO[†]

Abstract. This paper deals with image deblurring when the unknown original image is not smooth and a priori bounds on its derivatives cannot be prescribed in the inversion algorithm. A significant class of such deblurring problems occurring in medical, industrial, military, astronomical, and environmental applications is shown to be equivalent to backwards-in-time continuation in a generalized diffusion equation that may involve fractional Laplacians. The *slow-evolution-from-the-continuation-boundary* (SECB) constraint, introduced by the author in [*SIAM J. Numer. Anal.*, 31 (1994), pp. 1535–1557], is applicable to such nonsmooth image deblurring. A new analytical approach based on Fourier analysis provides sharp error estimates for SECB deblurring explicitly in terms of the constants entering the a priori constraints. It also leads to an explicit formula that expresses SECB's improvement over the classical Tikhonov–Miller method. An example from positron emission tomography (PET) imaging is used to illustrate the meaning of the SECB constraint. In this application, use of the SECB constraint reduces the L^2 norm of the Tikhonov–Miller inverse operator by almost a factor of ten.

Key words. ill-posed problems, infinitely smoothing operators, image deblurring, SECB restoration, Tikhonov–Miller restoration, nonsmooth images, PET imaging, error bounds

AMS subject classifications. 35R25, 35B60, 35B35, 65M30, 60E07, 68U10

PII. S0036141095290215

1. Introduction.

Explicit a priori bounds on derivatives of the unknown solutions underlie much of the analysis and numerical computation of ill-posed inverse problems in partial differential and integral equations. Mathematically, such a priori smoothness implies that solutions lie in a compact set in function space. Together with uniqueness of solutions, this leads to continuity of the inverse operator and stable numerical algorithms. On the physical side, smoothness constraints are appropriate in many problems of practical interest where solutions are known to possess relatively simple structures.

However, there are important classes of problems, notably in medical imaging, where solutions exhibit fairly complex structures, are typically not smooth, and often display singularities that may be of vital significance. In such problems, reconstruction methods that incorporate a priori smoothness assumptions on the solution are ill advised. The use of such procedures may result in visually pleasing but oversmoothed solutions in which significant diagnostic information has been eliminated.

In the absence of smoothness, error estimates that imply a *rate of convergence* as the data noise level $\epsilon \downarrow 0$ are not possible. However, error bounds valid for *fixed* small $\epsilon > 0$ that remain useful at realistic values of ϵ are possible, even when smoothness constraints are inapplicable. This was demonstrated in a recent paper [3] dealing with ill-posed continuation problems in partial differential equations. A new type of a priori constraint on solutions is introduced in [3], the so-called *slow-evolution-from-the-continuation-boundary* (SECB) constraint. The SECB constraint stabilizes

[*] Received by the editors August 11, 1995; accepted for publication (in revised form) January 26, 1996. A United States patent has been issued on part of the work described herein. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sima/28-3/29021.html

[†] Computing and Applied Mathematics Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899 (alfred@cam.nist.gov).

the inversion process against noise in the data, even when smoothness constraints are inapplicable. Applications include the harmonic continuation problem in the unit disc in the L^∞ norm, the spatial continuation problem for the heat equation in the L^2 norm, and the backwards-in-time continuation problem for self-adjoint parabolic equations in the L^2 norm. These three problems are canonical examples of ill-posed problems involving infinitely smoothing operators [6], [15].

This self-contained paper focuses on L^2 error bounds in *image deblurring*. An important class of image deblurring problems is shown to be equivalent to backwards-in-time continuation in a generalized diffusion equation involving fractional Laplacians. The SECB constraint can be applied to this ill-posed continuation problem. We bypass the Banach space approach used to obtain error bounds in [3] and rely instead on Fourier analysis in L^2. This method provides new and sharper estimates explicitly in terms of the constants entering the a priori constraints. It also leads to an explicit formula that expresses SECB's improvement over the classical Tikhonov–Miller method [16]. An example of a nonsmooth image from nuclear medicine illustrates the meaning of the SECB constraint and provides typical representative values for the constraint constants. This enables us to relate our analysis to real applications. Since the primary emphasis is on error bounds for the SECB method, no deblurring experiments are presented in this paper. However, such experiments are reported in [3, section 5], where the SECB approach is compared with three other image deblurring methods, using optimal values for the restoration parameters. These experiments confirm the analysis in the present paper by showing that the SECB constraint sharply reduces noise contamination.

2. Image deblurring with class-G point-spread functions.

We study a class of image restoration problems whereby the original image is reconstructed from a noisy blurred version [1], [7], [14], [21]. Following [21, Chapter 12], we consider space-invariant point-spread functions $p(x,y)$, and formulate the deblurring problem as the problem of solving the integral equation $Pg = f$, where

$$Pg \equiv \int_{R^2} p(x-u, y-v) g(u,v) du dv = f_e(x,y) + n(x,y) \equiv f(x,y). \tag{1}$$

Here $g(x,y)$ is the desired unblurred image, $f_e(x,y)$ is the blurred image that would have been recorded in the absence of noise, and $n(x,y)$ represents the cumulative effects of all noise processes that affect the final acquisition of the actual recorded image $f(x,y)$. This includes the case of *multiplicative noise*, where $n(x,y)$ is a function of $f_e(x,y)$. The noise component $n(x,y)$ is unknown but may be presumed small. Likewise, $f_e(x,y)$ is unknown. The type and intensity of the blurring caused by $p(x,y)$, together with the magnitude of $n(x,y)$, ultimately limit the quality of the restoration that can be achieved.

We denote by G the class of blurring kernels $p(x,y)$ with Fourier transform given by

$$\begin{aligned} \hat{p}(\xi,\eta) &\equiv \int_{R^2} p(x,y) e^{-2\pi i(\xi x + \eta y)} dx dy \\ &= e^{-\sum_{i=1}^{J} \alpha_i (\xi^2+\eta^2)^{\beta_i}}, \quad \alpha_i \geq 0, \quad 0 < \beta_i \leq 1. \end{aligned} \tag{2}$$

While many imaging phenomena are not described by (2), the latter encompasses highly significant applications. Thus if all α_is except one are zero, (2) reduces to

$\hat{p}(\xi,\eta) = e^{-\alpha(\xi^2+\eta^2)^\beta}$, which is the characteristic function of a Lévy "stable" probability distribution [5]. The case $\beta = 1$ corresponds to the Gaussian distribution and occurs in quite diverse contexts, including undersea imaging [27], nuclear medicine [20], [22], magnetic resonance imaging [17], computed tomography scanners [18], onboard optical seekers in cruise missiles [2], and ultrasonic imaging in nondestructive evaluation [12]. The case $\beta = 5/6$ describes long-exposure atmospheric-turbulence blurring [26]. The case $\beta = 1/2$ corresponds to the Cauchy distribution and has been used to model X-ray scattering in radiology [25]. Values of β satisfying $1/2 \leq \beta \leq 1$ characterize a wide variety of electron-optical devices [10], [11]. Such devices constitute important components in night-vision and undersea imaging systems [4], [11]. Modern biomedical imaging modalities such as II-TV fluoroscopic systems [23], selenium imaging plates [19], digital TV tomography systems [24], and radiographic screen-film systems [8], [13] are also based on electron-optical components. Methods exist for determining the values of α and β in each component electron-optical device [10]. In a typical imaging situation, several such components are commonly cascaded and used to image objects through a distorting medium such as the atmosphere or the ocean. The overall optical-transfer function is then given by (2), which is an example of an *infinitely divisible* characteristic function [5]. In many other applications, the general functional form described by (2) can be used to best-fit empirically determined optical-transfer functions by suitable choices of the parameters α_i, β_i, and J. In summary, the class G defined by (2) is worthy of analytical interest.

When the kernel of the integral operator P in (1) satisfies (2), the blurred noiseless image $f_e(x,y)$ may be identified with $u(x,y,1)$, where $u(x,y,t)$ is the unique bounded solution of the well-posed direct problem

$$\begin{aligned} &u_t = -\sum_{i=1}^{J} \gamma_i(-\Delta)^{\beta_i} u, \quad t > 0, \quad x,y \in R^2, \quad \gamma_i = \alpha_i(4\pi^2)^{-\beta_i}, \\ &u(x,y,0) = g(x,y), \end{aligned} \tag{3}$$

and $g(x,y)$ is the original unblurred image. This follows by recognizing (2) as the Fourier transform of the fundamental solution for the initial value problem (3). If all $\beta_i = 1$, (3) is the classical heat-conduction equation. For $0 < \beta_i \leq 1$, (3) represents a generalized diffusion process. The image restoration problem (1) is thus equivalent to backwards-in-time continuation of the solution of (3) from noisy data $f(x,y)$ at $t = 1$ rather than $f_e(x,y)$, where $f(x,y)$ is the blurred recorded image. The unblurred image $g(x,y)$ is the continuation at $t = 0$. For given fixed t with $0 < t < 1$, the continuation at time t represents a *partial restoration.* As $t \downarrow 0$, the partial restorations become sharper but noisier. Displaying a sequence of partial restorations at t_n as $t_n \downarrow 0$ can be helpful in identifying features which become obscured by noise at $t = 0$.

Fractional powers of the integral operator P in (1) are naturally related to the evolution equation (3). For fixed t with $0 \leq t \leq 1$, define P^t to be the convolution integral operator in $L^2(R^2)$, with kernel $p(t; x-u, y-v)$, where $p(t;x,y)$ is the inverse Fourier transform of $\exp\{-\sum_{i=1}^{J} t\alpha_i(\xi^2+\eta^2)^{\beta_i}\}$. Then $P^tP^s = P^{t+s}$ for all $s,t \geq 0$, and $P^0 = I$. The solution of (3) may be written as $u(t) = P^t g,\ 0 \leq t \leq 1$. Let $\|\ \|$ denote the $L^2(R^2)$ norm. By continuity, $\|P^t g - g\|$ is small for sufficiently small $t > 0$. Note that P is an infinitely smoothing operator.

Remark 1. The smallness of $\|P^s g - g\|$ for a given fixed $s > 0$ need not imply *any* smoothness in g. Fix $g \in L^2$, fix $s > 0$, and fix $\delta > 0$. If the α_i's are sufficiently

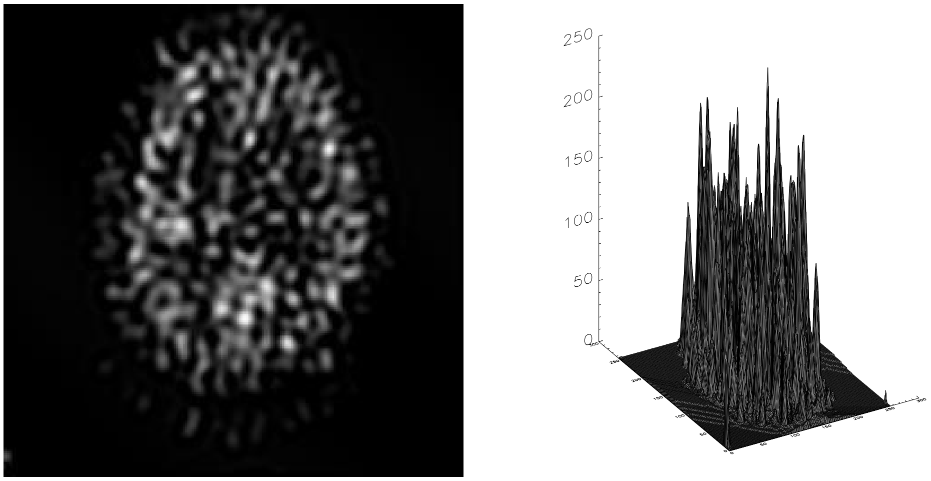


FIG. 1. *Left: PET image of transverse slice of human brain. Right: Image intensity as function of position $g(x, y)$.*

small, then

$$\int_{R^2} |\hat{g}(\xi,\eta)|^2 \left|1 - e^{-\sum_{i=1}^{J} s\alpha_i(\xi^2+\eta^2)^{\beta_i}}\right|^2 d\xi d\eta < \delta, \tag{4}$$

even if for every $q > 0$,

$$\int_{R2} (\xi^2 + \eta^2)^q |\hat{g}(\xi,\eta)|^2 d\xi d\eta = \infty. \tag{5}$$

As we will see below, the significance of the SECB constraint derives from this observation.

3. An example from nuclear medicine. The following example from positron emission tomography (PET) is helpful in motivating subsequent developments. Neuropsychiatrists have long been interested in correlating brain activity with such disorders as alcoholism, schizophrenia, dementia, Alzheimer's disease, mood disorders, and the like. PET imaging is a widely used modality in this field of research. A positron-emitting radionuclide is used to tag glucose molecules in their course through the brain. The metabolic rate of glucose is a key parameter that measures cerebral function and reflects the extent to which regions of the brain are working or failing to work. An emitted positron travels approximately 1 mm to 2 mm before colliding with an electron, resulting in an annihilating reaction in which two gamma ray photons are emitted at 180 degrees from each other. An extracranial ring of coincidence detectors, programmed to count emissions that are paired at 180 degrees, records every such positron emission. Following a coincidence, the source position can be mathematically reconstructed. This leads to an image of the distribution of the glucose tracer, which enables the radiologist to pinpoint areas of abnormal brain activity or to determine the health of cells.

A 256 × 256 PET image of a transverse slice of a human brain was obtained from the Nuclear Medicine Branch of the National Institutes of Health. That image,

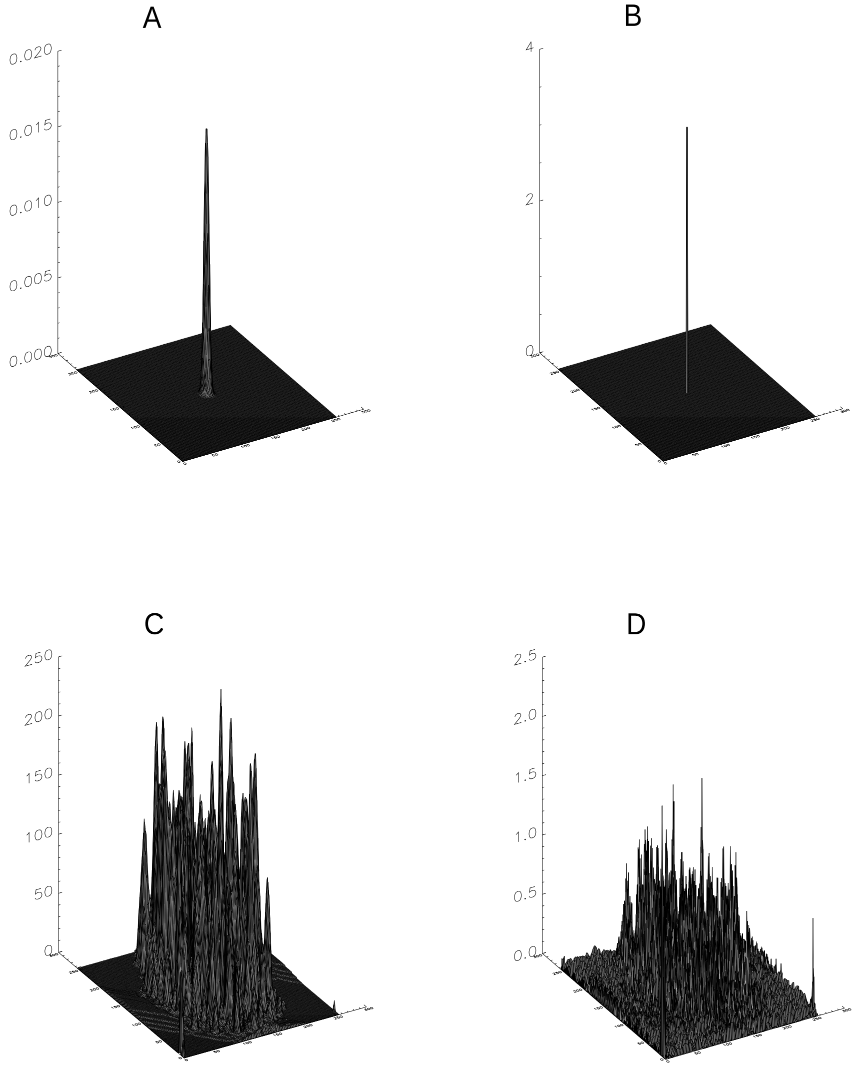


FIG. 2. (A) *The empirically determined PET-scanner Gaussian point-spread function* $p(x, y)$ *and the kernel of integral operator* P; (B) $q(x, y)$*, the* 200*th convolution root of* $p(x, y)$ *and the kernel of* $P^{0.005}$; (C) $P^{0.005}g$ *with* $g(x, y)$ *as in Fig.* 1; (D) $P^{0.005}g - g$. *The maximum intensity in* (B) *is* 200 *times that in* (A)*, while the maximum intensity in* (C) *is* 167 *times that in* (D).

displayed on the left in Fig. 1, consists of pixel values ranging from 0 to 255. The bright spots in the image represent areas of high positron emission and therefore high levels of brain activity. When the same pixel values are plotted as a function of x and y, we obtain the function $g(x, y)$ shown on the right in Fig. 1. Evidently, $g(x, y)$ is not a smooth function. Define the discrete L^2 norm of g by

$$\|g\| = \left\{ (256)^{-2} \sum_{j,k=1}^{256} g(x_j, y_k)^2 \right\}^{1/2}. \tag{6}$$

We find that $\|g\| = 40.15$. In general, owing to scattering of positrons prior to annihilation and to detector effects, reconstructed PET images are seldom as sharp as $g(x,y)$ shown in Fig. 1. Rather, a blurred image $f = Pg$ is obtained, where P is the convolution integral operator in (1), while $g(x,y)$ in Fig. 1 is the unknown original image in the deconvolution problem $Pg = f$. Some concentrated areas of moderate to high brain activity are quite often not apparent in the blurred image $f(x,y)$ but become evident only after deconvolution. Thus the mathematical problem becomes one of recovering g in Fig. 1 from some noisy blurred version f. The point-spread function associated with a given PET scanner can be determined experimentally by imaging a known point source. Typically, this empirically determined point-spread function is found to be well approximated by a Gaussian.

An example of a PET scanner Gaussian point-spread function $p(x,y)$ is shown in Fig. 2A. In Fig. 2B, $q(x,y)$, the 200th convolution root of $p(x,y)$ is shown. The function $q(x,y)$ is itself a Gaussian, one where the full width at half maximum is $\sqrt{200}$ times smaller than in $p(x,y)$. At the same time, the maximum value in $q(x,y)$ is 200 times larger than in $p(x,y)$. If P denotes the convolution operator with kernel $p(x,y)$, then $P^{0.005}$ is the operator with kernel $q(x,y)$. The function $q(x,y)$ is a good approximation of the Dirac δ-function in the following sense. The lack of smoothness of $g(x,y)$ in Fig. 1 notwithstanding, $P^{0.005}g$, shown in Fig. 2C, is almost identical to g. The difference, $(P^{0.005}g - g)$, is shown in Fig. 2D. In fact, we find that $\|P^{0.005}g - g\| = 0.123 = 0.003\|g\|$.

Let $\epsilon = 0.001\|g\| = 0.04$ represent an upper bound for the L^2 norm of the *noise* in the blurred image $f = Pg$. Smoothness constraints on the unknown sharp image g are not possible in the ill-posed deconvolution problem $Pg = f$. Instead, we have found that g satisfies the following *slow-evolution* constraint:

$$\|P^{0.005}g - g\| \le K\epsilon, \quad K \ge 3.1. \tag{7}$$

In section 6, we shall see how this constraint stabilizes the deconvolution problem $Pg = f$.

4. Error bounds in Tikhonov–Miller restoration. One of the best known techniques for regularizing ill-posed integral equations is the Tikhonov–Miller method [16]. In image deblurring [14], Tikhonov–Miller restoration is considered a canonical method. Seemingly more elaborate stochastic restoration procedures, such as Wiener filtering or maximum a posteriori (MAP) restoration [1], [14], ultimately result in mathematically similar expressions for the deblurred image. In addition, the Tikhonov–Miller method requires no a priori assumptions regarding the statistical character of the data noise.

In its simplest form, Tikhonov–Miller restoration requires an a priori bound ϵ for the L^2 norm of the *noise* in the blurred image f, together with an a priori bound M for the L^2 norm of the unblurred image g:

$$\|Pg - f\| \le \epsilon, \quad \|g\| \le M. \tag{8}$$

Here $\epsilon/M \ll 1$, and $\| \ \|$ denotes the $L^2(R^2)$ norm. It is assumed that ϵ and M are compatible with the existence of a $g(x,y) \in L^2$ satisfying (8). Tikhonov–Miller restoration [16] is defined as that unique function $g^T(x,y)$ which minimizes the functional

$$\|Ph - f\|^2 + \omega^2\|h\|^2, \quad \omega = \epsilon/M, \tag{9}$$

over all $h \in L^2(R^2)$. We have

$$g^T = [P^*P + \omega^2 I]^{-1}P^*f, \quad \omega = \epsilon/M. \tag{10}$$

Because no a priori bounds on derivatives of the unknown solution g were incorporated in (9), the best possible bound on the error in g^T is [16]

$$\|g - g^T\| \leq \sqrt{2}M, \tag{11}$$

irrespective of how small ϵ may be. See Remark 1 following Theorem 1 in [3]. The estimate in (11) cannot guarantee accuracy; in fact, appreciable noise contamination of the restored image is commonly experienced [3, section 5]. To get an error bound implying convergence as $\epsilon \downarrow 0$ in the Tikhonov–Miller method, stronger constraints need to be imposed on $g(x, y)$. Let L be an elliptic differential operator, let the unblurred image $g(x, y)$ be sufficiently smooth, and let an a priori bound $\|Lg\| \leq N$ be known. In this case, the appropriate Tikhonov–Miller functional to be minimized over all h in $L^2(R^2)$ is given by

$$\|Ph - f\|^2 + (\epsilon^2/N^2)\|Lh\|^2. \tag{12}$$

As shown in [6] and [15], an error bound of the form

$$\|g - g^T\| = O(N\{\log(N/\epsilon)\}^{-q}) \quad \text{as } \epsilon \downarrow 0 \tag{13}$$

is the best possible for (12). The value of $q > 0$ in (13) depends on the order of L and may be less than 1 if L is a fractional power Laplacian, for example. A major difficulty in connection with (12) lies in inferring a priori the correct value for N when the unknown image $g(x, y)$ is suspected of having localized singular behavior; there is the danger of underestimating N by several orders of magnitude. Moreover, while (13) implies convergence as $\epsilon \downarrow 0$, in practice, ϵ is small but *fixed*, depending as it does on the various noise processes inherent in the given imaging system. The logarithmic continuity result in (13) requires ϵ to be unrealistically small before the error bound becomes useful. Thus if $N = 1000$ and $q = 1$, we have $N\{\log(N/\epsilon)\}^{-1} \leq 10$ if and only if $\epsilon \leq 10^{-40}$. If $q = 1/2$, the requirement on ϵ is dramatically more severe.

5. SECB restoration. Like the Tikhonov–Miller method, the SECB approach requires no a priori knowledge of the statistical character of the data noise, but it does require an a priori bound ϵ for the L^2 norm of the noise in the blurred image f, together with an a priori bound M for the L^2 norm of the unblurred image g:

$$\|Pg - f\| \leq \epsilon, \quad \|g\| \leq M, \quad \epsilon/M \ll 1. \tag{14}$$

In addition, $\|P^s g - g\|$ is required to be small for some fixed s, $0 < s < 1$. This is always true by continuity if s is sufficiently small. We obtain an additional constraint on the class of solutions by requiring that s be known and not *too* small. Specifically, let ϵ and M be as above, with $\epsilon/M \ll 1$. For given K with $0 < K \ll M/\epsilon$, let s^* be defined by

$$s^* = \log\{M/(M - K\epsilon)\}/\log(M/\epsilon). \tag{15}$$

The SECB constraint requires that there exist a constant K with $0 < K \ll M/\epsilon$ such that

$$\|P^s g - g\| \leq K\epsilon, \quad s \text{ fixed}, \quad s^* < s < 1. \tag{16}$$

It is desirable (see section 6.2) that both the blurring operator P and the unblurred image $g(x, y)$ be such that (16) holds with a small $K > 0$ and a relatively large $s < 1$

so that $s/s^* \gg 1$. A large value of s/s^* reflects a priori knowledge that the given blurred image $f(x,y)$ has evolved slowly from the unknown sharp image $g(x,y)$, but it does not imply smoothness of $g(x,y)$. See (4) and (5) in Remark 1 in section 2. The four parameters $\{\epsilon,\ M,\ K,\ s\}$ constitute the a priori information in the SECB method. SECB restoration is then defined as the unique function $g^{\min}(x,y)$ which minimizes the functional

$$\|Ph - f\|^2 + \omega^2\|h\|^2 + K^{-2}\|h - P^s h\|^2, \quad \omega = \epsilon/M, \tag{17}$$

over all $h \in L^2(R^2)$. Note that SECB restoration reduces to Tikhonov–Miller restoration if $s = 0$ or $K = \infty$ in (17). When applicable, the SECB constraint produces error bounds that can be useful at realistic values of ϵ.

Remark 2. In applying the SECB constraint to an image where ϵ and M are known, a useful strategy is to first fix a small positive value of s. Since the point-spread function $p(x,y)$ is known, Fourier analysis may be used to obtain the kernel of P^s and determine how well that kernel approximates the Dirac δ-function. Prior experience with similar images, along the lines depicted in Fig. 2, may be used to arrive at an estimate for $\|P^s g - g\|$ and hence $\|P^s g - g\|/\epsilon$. The constant K in (16) is an upper bound for the latter quantity.

THEOREM 1. *Let the unblurred image $g(x,y)$ satisfy constraints* (14) *and* (16). *Let $g^{\min}(x,y)$ be the function which minimizes* (17). *Let $\omega = \epsilon/M$, and let $Q = Q(K,\omega,s)$ be the positive definite self-adjoint operator in $L^2(R^2)$ given by*

$$Q = P^*P + \omega^2 I + K^{-2}(I - P^s)^*(I - P^s). \tag{18}$$

Then $g^{\min}$ is the unique solution of $Qg^{\min} = P^*f$, and $g^{\min}$ satisfies

$$\|Pg^{\min} - f\|^2 + \omega^2\|g^{\min}\|^2 + K^{-2}\|(I - P^s)g^{\min}\|^2 \le 3\epsilon^2, \tag{19}$$

$$\|P(g - g^{\min})\|^2 + \omega^2\|g - g^{\min}\|^2 + K^{-2}\|(I - P^s)(g - g^{\min})\|^2 \le 3\epsilon^2, \tag{20}$$

$$\|g - g^{\min}\| \le \epsilon\sqrt{3}\|Q^{-1/2}\|. \tag{21}$$

Proof. Let $\mathcal{H}$ denote the Hilbert-space direct sum $L^2(R^2) \bigoplus L^2(R^2) \bigoplus L^2(R^2)$ with elements $[u,v,w]$, scalar product $([u_1,v_1,w_1],[u_2,v_2,w_2]) \equiv \langle u_1,u_2\rangle + \langle v_1,v_2\rangle + \langle w_1,w_2\rangle$, and norm $|||\ \ |||$. Let $\widetilde{P} : L^2(R^2) \mapsto \mathcal{H}$ be defined by $\widetilde{P}h = \big[Ph, \omega h, K^{-1}(I - P^s)h\big]$, and let $\widetilde{f} = [f,0,0]$. We seek to minimize $|||\widetilde{P}h - \widetilde{f}|||$ over all $h \in L^2$. The normal equation $\widetilde{P}^*\widetilde{P}g^{\min} = \widetilde{P}^*\widetilde{f}$ gives $Qg^{\min} = P^*f$, with Q as in (18). By hypothesis, $|||\widetilde{P}g - \widetilde{f}|||^2 \le 3\epsilon^2$. The minimizing element $g^{\min}$ is such that $\widetilde{P}g^{\min}$ is the orthogonal projection in $\mathcal{H}$ of $\widetilde{f}$ on the range of $\widetilde{P}$. By the Pythagorean theorem,

$$|||\widetilde{P}g^{\min} - \widetilde{f}|||^2 + |||\widetilde{P}(g - g^{\min})|||^2 = |||\widetilde{P}g - \widetilde{f}|||^2 \le 3\epsilon^2. \tag{22}$$

This gives (19) and (20). We now establish (21). From (18) and (20), we have

$$\|Q^{1/2}(g - g^{\min})\|^2 = \langle Q(g - g^{\min}), (g - g^{\min})\rangle = |||\widetilde{P}(g - g^{\min})|||^2 \le 3\epsilon^2. \tag{23}$$

Hence

$$\begin{aligned} \|g - g^{\min}\| &= \|Q^{-1/2}\ Q^{1/2}(g - g^{\min})\| \\ &\le \|Q^{-1/2}\|\|Q^{1/2}(g - g^{\min})\| \\ &\le \epsilon\sqrt{3}\|Q^{-1/2}\|. \end{aligned} \tag{24}$$

This completes the proof. ▯

6. Fourier analysis of SECB restoration. Inequality (21) in Theorem 1 reduces the problem of obtaining an error bound in SECB restoration to that of estimating $\|Q^{-1/2}(\epsilon, M, K, s)\|$. With $\hat{p}(\xi,\eta)$ as in (2), we have the following upon Fourier transforming $Q^{-1/2}h$:

$$(\widehat{Q^{-1/2}h})(\xi,\eta) = \left\{|\hat{p}(\xi,\eta)|^2 + \omega^2 + K^{-2}|1-\hat{p}^s(\xi,\eta)|^2\right\}^{-1/2}\hat{h}(\xi,\eta). \tag{25}$$

Using Parseval's theorem, $\|Q^{-1/2}h\| = \|(\widehat{Q^{-1/2}h})\|$, from which it follows that

$$\begin{aligned}\|Q^{-1/2}\| &= \sup_{\xi,\eta}\left\{|\hat{p}(\xi,\eta)|^2 + \omega^2 + K^{-2}|1-\hat{p}^s(\xi,\eta)|^2\right\}^{-1/2} \\ &= \left[\inf_{x\geq 0}\{e^{-2x}+\omega^2+K^{-2}(1-e^{-sx})^2\}\right]^{-1/2}, \quad \omega = \epsilon/M. \qquad (26)\end{aligned}$$

We now turn our attention to minimizing $\phi(x)$ on $x \geq 0$, where

$$\phi(x) = e^{-2x} + \omega^2 + K^{-2}(1-e^{-sx})^2, \quad \omega = \epsilon/M. \tag{27}$$

If $s = 0$, then $\phi(\infty) = \epsilon^2/M^2$ and $\|Q^{-1/2}\| = M/\epsilon$. In this case, we recover the Tikhonov–Miller error bound (11). More generally, the next lemma shows why we need $s > s^*$ in the SECB constraint (16).

LEMMA 1. *Let s^* be as in* (15). *If $s \leq s^*$, $M/(\epsilon\sqrt{3}) \leq \|Q^{-1/2}\| \leq (M/\epsilon)$.*

Proof. Let $x_0 = \log(M/\epsilon)$. Then $e^{-2x_0} = \omega^2$ and $1 - e^{-sx_0} \leq K\omega$ if $s \leq s^*$. Hence $\omega^2 \leq \phi(x_0) \leq 3\omega^2$ if $s \leq s^*$, and the result follows from (26). □

LEMMA 2. *Let $s > 0$. Then on $x \geq 0$, $\phi(x)$ has a unique minimum at $x = \bar{x} > 0$ where*

$$e^{-2\bar{x}} = sK^{-2}(e^{-s\bar{x}} - e^{-2s\bar{x}}). \tag{28}$$

Moreover, if $K/s > 1$,

$$\frac{2e}{1+2e}\log\left\{\frac{K}{s}\right\} \leq \bar{x} \leq \frac{1}{2-s}\log\left\{\frac{K^2}{s\left[1-(s/K)^{2es/(1+2e)}\right]}\right\}. \tag{29}$$

Proof. We have $\phi'(x) = 0$ if and only if $e^{-2x} = sK^{-2}(e^{-sx} - e^{-2sx})$. Let $\bar{x}$ be the abscissa of the unique point where the curve $y = e^{-2x}$ intersects the curve $y = sK^{-2}(e^{-sx} - e^{-2sx})$. We have $\phi''(\bar{x}) = 4e^{-2\bar{x}} - 2s^2K^{-2}e^{-s\bar{x}} + 4s^2K^{-2}e^{-2s\bar{x}}$. Using $2s^2K^{-2}e^{-s\bar{x}} = 2se^{-s\bar{x}} + 2s^2K^{-2}e^{-2s\bar{x}}$, we find that $\phi''(\bar{x}) = (4-2s)e^{-2\bar{x}} + 2s^2K^{-2}e^{-2s\bar{x}}$. Since $0 < s \leq 1$, $\phi''(\bar{x}) > 0$. Thus $\bar{x}$ is the unique minimum. Next, $e^{-2\bar{x}} = s^2K^{-2}\bar{x}(e^{-s\bar{x}} - e^{-2s\bar{x}})/s\bar{x} \leq s^2K^{-2}\bar{x}$ since $f(y) \equiv (e^{-y} - e^{-2y})/y$ is a monotone decreasing function with a maximum value of 1 at $y = 0$. If $0 < \bar{x} < 1$, then $e^{-2\bar{x}} < s^2K^{-2}$, which implies that $\bar{x} > \log(K/s)$. If $\bar{x} \geq 1$, then $\log(\bar{x})/\bar{x} \leq 1/e$ and $1 + \log(\bar{x})/2\bar{x} \leq (1+2e)/2e$. From $\bar{x}e^{2\bar{x}} \geq K^2/s^2$, we get $\bar{x}\{1 + \log(\bar{x})/2\bar{x}\} \geq \log(K/s)$. Hence $\bar{x} \geq \{2e/(1+2e)\}\log(K/s)$ if $\bar{x} \geq 1$, and the inequality remains valid if $0 < \bar{x} < 1$. To obtain the upper bound on $\bar{x}$ in (29), first observe that $1 - e^{-s\bar{x}} \geq 1 - \{s/K\}^{2es/(1+2e)}$. Hence from (28), $e^{-(2-s)\bar{x}} = sK^{-2}(1 - e^{-s\bar{x}}) \geq sK^{-2}\left[1 - \{s/K\}^{2es/(1+2e)}\right]$. The result follows upon taking logarithms. □

LEMMA 3. *If $0 < K/s \leq 2$, $\|Q^{-1/2}\| < 5$.*

Proof. From (28), $\bar{x}e^{2\bar{x}} = (K^2s^{-2})\{s\bar{x}/(e^{-s\bar{x}} - e^{-2s\bar{x}})\} \leq 4s\bar{x}/(e^{-s\bar{x}} - e^{-2s\bar{x}})$ if $K/s \leq 2$. Hence $(1/4)\bar{x}e^{(2-s)\bar{x}} \leq s\bar{x}/(1 - e^{-s\bar{x}}) \leq \bar{x}/(1 - e^{-\bar{x}})$, the last inequality resulting from the fact that $g(y) \equiv y/(1-e^{-y})$ is a monotone increasing function on $y \geq 0$, and $0 < s \leq 1$. Therefore, $e^{\bar{x}} \leq e^{(2-s)\bar{x}} \leq 4(1-e^{-\bar{x}})^{-1}$. Thus $\bar{x} \leq \log 5$, and $e^{-2\bar{x}} \geq 1/25$. The result now follows from (26). Note that from (15), $0 < K/s \leq 2$ implies $s/s^* \geq \{(M-2s\epsilon)/(2\epsilon)\}\log(M/\epsilon) \gg 1$. □

LEMMA 4. *Let $K/s > 1$, and let $x_0 \geq (1/(2-s))\log\left[K^2/s\left\{1-(s/K)^{2es/(1+2e)}\right\}\right]$. Consider the iteration*

$$x_{n+1}e^{2x_{n+1}} = K^2 s^{-2}\left\{sx_n/(e^{-sx_n} - e^{-2sx_n})\right\}, \quad n = 0, 1, 2, \ldots. \tag{30}$$

Then $0 < x_{n+1} \leq x_n \leq \cdots \leq x_1 \leq x_0$, and the sequence $\{x_n\}$ converges to $\bar{x}$.

Proof. Let $h(y) \equiv y/(e^{-y} - e^{-2y})$. Then $h(y)$ and ye^{2y} are monotone increasing functions on $y \geq 0$. Since $0 < s \leq 1$, the function xe^{2x} eventually increases faster than $K^2s^{-2}h(sx)$, and the two curves intersect at $x = \bar{x}$. In particular, $xe^{2x} \geq K^2s^{-2}h(sx)$ if $x \geq \bar{x}$. Therefore, using (29) in Lemma 2, $x_1e^{2x_1} = K^2s^{-2}h(sx_0) \leq x_0e^{2x_0}$, which implies that $x_1 \leq x_0$. Let $A_n = K^2s^{-2}h(sx_n)$ so that $x_{n+1}e^{2x_{n+1}} = A_n$. Then $x_{m+1} \leq x_m$ implies $A_{m+1} \leq A_m$. It follows that $0 < x_{n+1} \leq x_n \leq \cdots \leq x_1 \leq x_0$, and $K^2s^{-2} < A_{n+1} \leq A_n \leq \cdots \leq A_1 \leq A_0$. Therefore, x_n converges to $z > 0$, and A_n converges to $K^2s^{-2}h(sz)$, which implies $z = \bar{x}$. □

THEOREM 2. *Let $s > s^*$, let g satisfy* (14) *and* (16), *and let $g^{\min}$ be as in Theorem* 1. *If $0 < K/s \leq 2$, then $\|g - g^{\min}\| < 5\epsilon\sqrt{3}$. If $K/s > 2$, let*

$$\begin{aligned} A &= \left\{sK^{-2}\left(1 - \{s/K\}^{2es/(1+2e)}\right)\right\}^{2/(2-s)}, \\ B &= \{\epsilon/M\}^2, \\ C &= K^{-2}\left\{1 - (s/K)^{2es/(1+2e)}\right\}^2. \end{aligned} \tag{31}$$

Then

$$\|Q^{-1/2}\| \leq \{A + B + C\}^{-1/2} \tag{32}$$

and

$$\|g - g^{\min}\| \leq \epsilon\sqrt{3}\,\{A + B + C\}^{-1/2}. \tag{33}$$

Proof. If $0 < K/s \leq 2$, the result follows from Lemma 3 and (21). If $K/s > 2$, we can use the upper and lower bounds for $\bar{x}$ in Lemma 2 to estimate $\phi(\bar{x})$. We find $\phi(\bar{x}) \geq A + B + C$, where A, B, and C are as in (31). The result follows from (21) together with $\|Q^{-1/2}\| = \{\phi(\bar{x})\}^{-1/2}$. □

6.1. An example. The above analysis produces reliable error bounds in SECB restoration. In the PET imaging example of section 3, we have $M = 40.15$, $\epsilon = 0.04$, $s = 0.005$, and $K = 3.1$. This gives $K/s = 620$, $s^* = 4.47 \times 10^{-4}$, and $s/s^* = 11.2$. Lemma 2 provides upper and lower bounds for $\bar{x}$ for given K and s. From (29), we get $5.4308 \leq \bar{x} \leq 5.6045$. From (31) in Theorem 2, given ϵ, M, K, and s, we can find an upper bound for $\|Q^{-1/2}\| = \{\phi(\bar{x})\}^{-1/2} \leq \{A + B + C\}^{-1/2}$. We get $\|Q^{-1/2}\| \leq 105.87$. However, we may also use the iteration in Lemma 4 to calculate $\bar{x}$ to high accuracy. Evaluating $\phi(\bar{x})$ then provides the *exact* value for $\|Q^{-1/2}\|$ as in (26). We find $\bar{x} = 5.5902$ and $\{\phi(\bar{x})\}^{-1/2} = \|Q^{-1/2}\| = 103.14$. Evidently, the estimates provided by Lemma 2 and Theorem 2 are in close agreement with the exact values in this example.

As previously noted, SECB restoration reduces to Tikhonov–Miller restoration ((9)), if $s = 0$. The exact value for $\|Q^{-1/2}\|$ when $s = 0$ is $M/\epsilon = 1004$. The present value, $\|Q^{-1/2}\| = 103$, is about ten times smaller. This illustrates the stabilizing property of the SECB constraint. We obtain the estimate $\|g - g^{\min}\| \leq 7.15$ for the SECB image, versus $\|g - g^T\| \leq 56.78$ for the Tikhonov–Miller image.

In [3, section 2], a one-dimensional SECB deblurring example is given where $\epsilon = 10^{-3}$, $M = 10$, $K = 3$, and $s = 0.01$. In this case, $s/s^* = 307$, and $\|Q^{-1/2}\| = 56.86$. Without the SECB constraint, $\|Q^{-1/2}\| = M/\epsilon = 10^4$, which is 175 times larger.

6.2. Improvement over Tikhonov–Miller and the ratio (s/s^*). It follows from (27) that $\inf_{x \geq 0} \phi(x) > \epsilon^2/M^2$ whenever $s > 0$. Hence from (26), the SECB bound for $\|Q^{-1/2}\|$ is *always* smaller than the corresponding Tikhonov–Miller bound M/ϵ. However, as may be surmised from Lemma 1 and the above-mentioned examples, SECB's improvement over the Tikhonov–Miller value increases as the value of s/s^* increases. In the case where $K/s \leq 2$, this was noted at the end of the proof of Lemma 3. For the case where $K/s > 2$, the role played by s/s^* in the estimate of Theorem 2 may be discerned through the following analysis.

With A, B, and C as in (31), let $b = 2e/(1+2e) \approx 0.845$, and write $\{s/K\}^{bs} = \exp\{-bs\log(K/s)\} \approx 1 - bs\log(K/s)$ for small values of s such as typically enter the SECB constraint. We may likewise replace the exponent $2/(2-s)$ by unity. Then

$$(34) \qquad A \approx bs^2K^{-2}\log(K/s), \qquad C \approx b^2s^2K^{-2}\{\log(K/s)\}^2, \quad (K/s) > 2.$$

Therefore,

$$(35) \qquad (A+B+C)^{-1/2} \approx \frac{(K/s)}{[b^2\{\log(K/s)\}^2 + b\log(K/s) + \{K\epsilon/(sM)\}^2]^{1/2}}.$$

From (15), we have $s^* \approx K\epsilon/\{M\log(M/\epsilon)\}$ for $K\epsilon \ll M$. Hence

$$(36) \qquad K/s \approx (s^*/s)(M/\epsilon)\log(M/\epsilon).$$

The following result, which shows that $\|Q^{-1/2}\| \ll M/\epsilon$ whenever $s^*/s \ll 1$, is immediate from (32), (35), and (36).

THEOREM 3. *Let $K/s > 2$, let $b = 2e/(1+2e)$, and let A, B, and C be as in* (31). *For small $s > 0$, we have*

$$\|Q^{-1/2}\| \leq (A+B+C)^{-1/2}$$

$$(37) \qquad \approx \frac{(s^*/s)(M/\epsilon)\log(M/\epsilon)}{[b^2\{\log(K/s)\}^2 + b\log(K/s) + \{K\epsilon/(sM)\}^2]^{1/2}}, \quad (K/s) > 2.$$

6.3. Behavior as $\epsilon \downarrow$ o. While the SECB constraint can result in useful error bounds at realistic values of ϵ, the estimates in Theorems 2 and 3 do not imply convergence as $\epsilon \downarrow 0$. With s fixed in the SECB constraint (16), $\epsilon \downarrow 0$ implies $K \uparrow \infty$, and Lemma 2 then shows that $\bar{x} \uparrow \infty$. Thus $\phi(\bar{x}) \to 0$. However, $\epsilon\{\phi(\bar{x})\}^{-1/2}$ remains bounded, and $\|g - g^{\min}\|$ becomes small, on the order of $\|P^sg - g\|$, as $\epsilon \downarrow 0$. We have the following result.

THEOREM 4. *Let $\|g\| \leq M$. Fix $s > 0$, let $\sigma = \|P^sg - g\|$, and let $b = 2e/(1+2e)$. For each $\epsilon > 0$, let $K = \sigma/\epsilon$, let $\omega = \epsilon/M$, and let $g^{\min}$ be the unique solution of $Qg^{\min} = P^*f$, where $Q = P^*P + \omega^2I + K^{-2}(I-P^s)^*(I-P^s)$. Then*

$$(38) \qquad \|g - g^{\min}\| < \sqrt{3}\|P^sg - g\| + O\{(s\epsilon)^{sb}\} \quad \textit{as } \epsilon \downarrow 0.$$

Proof. With $K = \sigma/\epsilon$, we have from (27) that

$$\begin{aligned}\epsilon^{-2}\phi(\bar{x}) &= \epsilon^{-2}e^{-2\bar{x}} + M^{-2} + \sigma^{-2}(1-e^{-s\bar{x}})^2 \\ &> \sigma^{-2}(1-e^{-s\bar{x}})^2 > \sigma^{-2}(1-(s\epsilon/\sigma)^{sb})^2\end{aligned} \tag{39}$$

upon using (29) in Lemma 2. From Theorem 2, we have $\|g - g^{\min}\| \le \epsilon\sqrt{3}\{\phi(\bar{x})\}^{-1/2}$ for every $\epsilon > 0$. The result follows. □.

7. Summary. A priori information is essential in the analysis of ill-posed continuation problems in partial differential equations [9]. The classical approach to regularization is based on prescribed bounds on the derivatives of the unknown solution. This approach suffers from three major drawbacks. First, derivatives may fail to exist in many problems of practical interest. Second, when such derivatives exist, it may not be possible to estimate them reliably from a priori considerations. Third, when an estimate N on some Sobolev norm of the desired solution is known, the best possible error bound for the regularized solution has the form $O(N\{\log(N/\epsilon)\}^{-q})$, $q > 0$. This is the notorious *logarithmic continuity* described in Fritz John's writings on ill-posed problems. Here ϵ, the norm of the data noise, is fixed and small, but it is seldom small enough to render that error bound meaningful.

A different method of regularization was examined in this paper for continuation problems of the form $Pg = f$, where P is an infinitely smoothing integral operator with known kernel $p(x, y)$. This method does not require differentiability of the unknown solution g, yet it leads to useful error bounds at realistic values of ϵ. The method is based on the observation that in many applications where the unknown g may not be differentiable, both g and the integral operator P are sufficiently well-behaved that for some fixed small $s > 0$, P^s almost acts like the identity operator when applied to g. Thus $\|P^s g - g\|$ can be expected to be small. This idea was illustrated in Fig. 2 using an example from nuclear medicine. The SECB constraint consists of imposing the requirement that $\|P^s g - g\|/\epsilon$ be bounded by some known constant K. The constraint becomes effective when s can be chosen relatively large with K relatively small so that $s^*/s \ll 1$. In Theorem 2, error bounds were obtained for the SECB approach explicitly in terms of the constants ϵ, M, K, and s entering the a priori constraints. In Theorem 3, an explicit formula was derived that compares the L^2 norm of the SECB inverse operator to that of the Tikhonov–Miller inverse operator. The improvement was shown to be governed by the ratio s^*/s. The SECB error bound for the regularized solution may be several hundred times smaller than that for the Tikhonov–Miller method if s^*/s is sufficiently small. Image deblurring experiments reported in [3] show that substantial improvement occurs even for moderately small values of s^*/s.

REFERENCES

[1] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*, Prentice–Hall, Englewood Cliffs, NJ, 1977.

[2] M. Banish, R. Clark, and A. Kathman, *A validated code to predict the performance of onboard broadband optical seekers through a turbulent transonic flow*, AIAA 92-2792, in Proc. 1992 AIAA SDIO Annual Interceptor Technology Conference, American Institute of Aeronautics and Astronautics, Washington, DC, 1992.

[3] A. S. Carasso, *Overcoming Hölder continuity in ill-posed continuation problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1535–1557.

[4] K. A. Costello, V. W. Aebi, and H. F. Macmillan, *Imaging GaAs vacuum photodiode with 40% quantum efficiency at* 530 *nm*, SPIE Proc., 1243 (1990), pp. 99–106.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., John Wiley, New York, 1971.

[6] J. N. Franklin, *On Tikhonov's method for ill-posed problems*, Math. Comp., 28 (1974), pp. 889–907.

[7] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, 2nd ed., Addison–Wesley, Reading, MA, 1987.

[8] Y. Higashida, K. D. Doi, and H. B. MacMahon, *Dual-film cassette technique for studying the effect of radiographic image quality on diagnostic accuracy*, Med. Phys., 11 (1984), pp. 646–652.

[9] 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.

[10] C. B. Johnson, *Classification of electron-optical device modulation transfer function*, Adv. Electron. Electron Phys., 33B (1972), pp. 579–584.

[11] C. B. Johnson, S. B. Patton, and E. Bender, *High resolution microchannel plate image tube development*, SPIE Proc., 1449 (1991), pp. 2–12.

[12] P. Karpur and B. G. Frock, *Two-dimensional pseudo-Wiener filtering in ultrasonic imaging for nondestructive evaluation applications*, in Review of Progress in Quantitative Nondestructive Evaluation, Vol. 8A, Plenum Press, New York, 1989, pp. 743–750.

[13] H. Kuhn and W. Knupfer, *Imaging characteristics of different mammographic screens*, Med. Phys., 19 (1992), pp. 449–457.

[14] R. L. Lagendijk and J. Biemond, *Iterative Identification and Restoration of Images*, Kluwer Academic Publishers, Norwell, MA, 1991.

[15] B. A. Mair, *Tikhonov regularization for finitely and infinitely smoothing operators*, SIAM J. Math. Anal., 25 (1994), pp. 135–147.

[16] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.

[17] S. M. Mohapatra, J. D. Turley, J. R. Prince, J. C. Blechinger, and D. A. Wilson, *Transfer function measurement and analysis for a magnetic resonance imager*, Med. Phys., 18 (1991) pp. 1141–1144.

[18] E. L. Nickoloff and R. Riley, *A simplified approach for modulation transfer function determinations in computed tomography*, Med. Phys., 12 (1985), pp. 437–442.

[19] P. J. Papin and H. K. Huang, *A prototype amorphous selenium imaging plate*, Med. Phys., 14 (1987), pp. 322–329.

[20] K. S. Pentlow, M. C. Graham, R. M. Lambrecht, K. V. Cheung, and S. M. Larson, *Quantitative imaging of* I-124 *using positron emission tomography with applications to radioimmunodiagnosis and radioimmunotherapy*, Med. Phys., 18 (1991), pp. 357–366.

[21] W. K. Pratt, *Digital Image Processing*, 2nd ed., John Wiley, New York, 1991.

[22] U. Raff, D. N. Stroud, and W. R. Hendee, *Improvement of lesion detection in scintigraphic images by SVD techniques for resolution recovery*, IEEE Trans. Med. Imaging, MI-5 (1986), pp. 35-44.

[23] S. Rudin, D. Bednarek, and R. Wong, *Improving fluoroscopic image quality with continuously variable zoom magnification*, Med. Phys., 19 (1991), pp. 972–977.

[24] M. Takahashi, S. Yoshioka, H. Bussaka, Y. Higashida, M. Kamiya, and M. Tsuneoka, *Digital TV tomography: Description and physical assessment*, Med. Phys., 17 (1990), pp. 681–685.

[25] F. C. Wagner, A. Macovski, and D. G. Nishimura, *A characterization of scatter point spread functions in terms of air gaps*, IEEE Trans. Med. Imaging, 7 (1988), pp. 337–344.

[26] J. C. Wyant, ed., *Imaging through the atmosphere*, SPIE Proc., 75 (1976).

[27] H. T. Yura, *Imaging in clear ocean water*, Appl. Optics, 12 (1973), pp. 1061–1066.

SIAM J. Math. Anal.
Vol. 28, No. 3, pp. 669–703, May 1997

010

EXPONENTIAL ASYMPTOTICS IN A SINGULAR LIMIT FOR n-LEVEL SCATTERING SYSTEMS[*]

ALAIN JOYE[†]

Abstract. The singular limit $\varepsilon \to 0$ of the S-matrix associated with the equation $i\varepsilon d\psi(t)/dt = H(t)\psi(t)$ is considered, where the analytic generator $H(t) \in M_n(\mathbf{C})$ is such that its spectrum is real and nondegenerate for all $t \in \mathbf{R}$. Sufficient conditions allowing us to compute asymptotic formulas for the exponentially small off-diagonal elements of the S-matrix as $\varepsilon \to 0$ are made explicit and a wide class of generators for which these conditions are verified is defined. These generators are obtained by means of generators whose spectrum exhibits eigenvalue crossings which are perturbed in such a way that these crossings turn into avoided crossings. The exponentially small asymptotic formulas which are derived are shown to be valid up to exponentially small relative error by means of a joint application of the complex Wentzel–Kramers–Brillouin (WKB) method together with superasymptotic renormalization. This paper concludes with the application of these results to the study of quantum adiabatic transitions in the time-dependent Schrödinger equation and of the semiclassical scattering properties of the multichannel stationary Schrödinger equation. The results presented here are a generalization to n-level systems, $n \geq 2$, of results previously known for two-level systems only.

Key words. singular perturbations, semiclassical and adiabatic approximations, n-level S-matrix, turning-point theory

AMS subject classifications. 34E20, 34L25, 81Q20

PII. S0036141095288847

1. Introduction.

Several problems of mathematical physics lead to the study of the scattering properties of linear ordinary differential equations in a singular limit

$$i\varepsilon\psi'(t) = H(t)\psi(t), \quad t \in \mathbf{R}, \quad \varepsilon \to 0, \tag{1.1}$$

where the prime denotes the derivative with respect to t, $\psi(t) \in \mathbf{C}^n$, and $H(t) \in M_n(\mathbf{C})$ for all t. A system described by such an equation will be called an n-level system. Let us mention, for example, the study of the adiabatic limit of the time-dependent Schrödinger equation or the semiclassical limit of the one-dimensional multichannel stationary Schrödinger equation at energies above the potential barriers, to which we will return below. When the generator $H(t)$ is well behaved at $+\infty$ and $-\infty$, the scattering properties of the problem can be described by means of a matrix naturally associated with equation (1.1), the so-called S-matrix. This matrix relates the behavior of the solution $\psi(t)$ as $t \to -\infty$ to that of $\psi(t)$ as $t \to +\infty$. Assuming that the spectrum $\sigma(t)$ of $H(t)$ is real and nondegenerate,

$$\sigma(t) = \{e_1(t) < e_2(t) < \cdots < e_n(t)\} \in \mathbf{R}, \tag{1.2}$$

the S-matrix is essentially given by the identity matrix

$$S = \mathrm{diag}(s_{11}(\varepsilon), s_{22}(\varepsilon), \ldots, s_{nn}(\varepsilon)) + \mathcal{O}(\varepsilon^\infty), \quad \text{where } s_{jj}(\varepsilon) = 1 + \mathcal{O}(\varepsilon) \text{ as } \varepsilon \to 0, \tag{1.3}$$

[*] Received by the editors July 10, 1995; accepted for publication (in revised form) February 16, 1996.

http://www.siam.org/journals/sima/28-3/28884.html

[†] Centre de Physique Théorique, CNRS Marseille, Luminy Case 907, F-13288 Marseille cedex 9, France and Phymat, Université de Toulon et du Var, B.P. 132, 83957 La Garde cedex, France (joye@cpt.univ-mrs.fr).

provided $H(t)$ is C^∞; see, e.g., [F1], [F2], and [W]. Moreover, if $H(t)$ is assumed to be analytic, it was proven in various situations that the off-diagonal elements s_{jk} of S are exponentially decreasing [FF], [W], [F1], [F2], [JKP], [JP4]:

$$s_{jk} = \mathcal{O}\left(\mathrm{e}^{-\kappa/\varepsilon}\right), \quad \forall j \neq k, \tag{1.4}$$

as $\varepsilon \to 0$. See also [JP1], [N], [M], and [Sj] for corresponding results in infinite-dimensional spaces. Since the physical information is often contained in these off-diagonal elements, it is of interest to be able to give an asymptotic formula for s_{jk} rather than a mere estimate.

For two-level systems (or systems reducible to this case (see [JP2], [J], and [MN])), the situation is now reasonably well understood, at least under generic circumstances. Indeed, a rigorous study of the S-matrix associated with (1.1) when $n = 2$ under the hypotheses loosely stated above is provided in the recent paper [JP4]. The treatment presented unifies, in particular, earlier results obtained for either the time-dependent adiabatic Schrödinger equation (see, e.g., [JP3] and the references therein) or the study of the above barrier reflexion in the semiclassical limit (see, e.g., [FF] and [O]). Further references are provided in [JP4]. As an intermediate result, the asymptotic formula

$$s_{jk} = g_{jk}\mathrm{e}^{-\Gamma_{jk}/\varepsilon}\,(1 + \mathcal{O}(\varepsilon)), \quad \varepsilon \to 0, \tag{1.5}$$

for $j \neq k \in \{1, 2\}$ with $g_{jk} \in \mathbf{C}$ and $\mathrm{Re}\Gamma_{jk} > 0$ is proven in [JP4]. As is well known, to get an asymptotic formula for s_{jk}, one has to consider (1.1) in the complex plane, in particular in the vicinity of the degeneracy points of the analytic continuations of eigenvalues $e_1(z)$ and $e_2(z)$. Provided the level lines of the multivalued function

$$\mathrm{Im}\int_0^z e_1(z') - e_2(z')dz' = \mathrm{cst}, \tag{1.6}$$

called Stokes lines, naturally associated with (1.1) behave properly in the complex plane, the so-called complex Wentzel–Kramers–Brillouin (WKB) method allows to prove (1.5). More importantly, however, it is also shown in [JP4] how to improve (1.5) to an asymptotic formula accurate up to an exponentially small relative error:

$$s_{jk} = g^*_{jk}(\varepsilon)\mathrm{e}^{-\Gamma^*_{jk}(\varepsilon)/\varepsilon}(1 + \mathcal{O}(\mathrm{e}^{-\kappa/\varepsilon})), \quad \varepsilon \to 0, \tag{1.7}$$

with $g^*_{jk}(\varepsilon) = g_{jk} + \mathcal{O}(\varepsilon)$ and $\Gamma^*_{jk}(\varepsilon) = \Gamma_{jk} + \mathcal{O}(\varepsilon^2)$. This is achieved by using a complex WKB analysis jointly with the recently developed superasymptotic theory [Be], [N], [JP2]. Note that when given a generator, the principal difficulty in justifying formulas (1.5) and (1.7) is the verification that the corresponding Stokes lines (1.6) display the proper behavior *globally* in the complex plane, which may or may not be the case [JKP]. However, this condition is always satisfied when the complex eigenvalue degeneracy is close to the real axis, as shown in [J]. See also [MN] and [R] for recent related results.

For n-level systems, with $n \geq 3$, the situation is by no means as well understood. There are some results obtained with particular generators. In [D], [CH1], [CH2], and [BE], certain elements of the S-matrix are computed if $H(t) = H^*(t)$ depends linearly on t, $H(t) = A+tB$ for some particular matrices A and B. The choices of A and B are such that all components of the solution $\psi(t)$ can be deduced from the first one and an exact integral representation of this first component can be obtained. The integral representation is analyzed by standard asymptotic techniques, and this leads to results which are valid for any $\varepsilon > 0$, as in the case for the classical Landau–Zener generator. The study of the three-level problem when $H(t) = H^*(t) \in M_3(\mathbf{R})$ is tackled in the closing section of the very interesting paper [HP]. A nonrigorous and essentially local discussion of the behavior of the level lines of $\operatorname{Im} \int_0^z e_j(z') - e_k(z')dz'$, $j \neq k = 1, 2, 3$, is provided, and it justifies in very favorable cases an asymptotic formula for some elements of the S-matrix. See also the review [So], where a nonrigorous study of (1.1) is made close to a complex degeneracy point of a group of eigenvalues by means of an exact solution to a model equation. However, no asymptotic formula for s_{jk}, $j \neq k$, can be found in the literature for general n-level systems, $n \geq 3$. This is due to the fact that the direct generalization of the method used successfully for two-level systems may lead to seemingly inextricable difficulties for $n = 3$. Indeed, with three eigenvalues, one has to consider three sets of level lines $\operatorname{Im} \int_0^z e_j(z') - e_k(z')dz'$ to deal with (1.1) in the complex plane, and the conditions that they have to fulfill in order for the limit $\varepsilon \to 0$ to be controlled may be incompatible for a given generator; see [F1], [F2], and [HP]. It should be mentioned, however, that there are specific examples in which this difficult problem can be mastered. Such a result was recently obtained in the semiclassical study [Ba] of a particular problem of resonances for which similar considerations in the complex plane are required.

The goal of this paper is to provide some general insight into the asymptotic computation of the S-matrix associated with n-level systems, $n \geq 3$, based on a generalization of the techniques which proved to be successful for two-level systems. The content of this paper is twofold. On one hand, we set up a general framework in which asymptotic formulas for the exponentially small off-diagonal coefficients can be proven. On the other hand, we actually prove such formulas for a wide class of n-level systems. In the first part of the paper, we give our definition of the S-matrix associated with equation (1.1) and make explicit the symmetries it inherits from the symmetries of $H(t)$ for $t \in \mathbf{R}$ (Proposition 2.1). We then turn to the determination of the analyticity properties of the eigenvalues and eigenvectors of $H(z)$, $z \in \mathbf{C}$, which are at the root of the asymptotic formulas that we derive later (Lemma 3.1). The next step is the formulation of sufficient conditions adapted to the scattering situation that we consider, under which a complex WKB analysis allows us to prove a formula like (1.5) (Proposition 4.1). The conditions stated are similar but not identical to those given in [JKP] or [HP]. As a final step, we show how to improve the asymptotic formula (1.5) to (1.7) by means of superasymptotic machinery (Proposition 5.2 and Lemma 5.2). We then turn to the second part of the paper, where we show that a wide class of generators fits into our framework and satisfies our conditions. These generators are obtained by perturbation of generators whose eigenvalues display degeneracies on the real axis (in the spirit of [J]). We prove that for these generators, in the absence of any symmetry of the generator $H(t)$, at least one element per column in the S-matrix can be asymptotically computed (Theorem 6.1). This is the main technical section of the paper. The major advantage of this construction is that it is sufficient to look at the behavior of the eigenvalues on the real axis to check if the conditions are satisfied. The closing section contains an application of our general results to the study of quantum adiabatic transitions in the time-dependent Schrödinger equation and of the semiclassical scattering properties of the multichannel stationary Schrödinger equation. In particular, we make explicit use of the symmetries of the S-matrix to increase the number of its elements for which an asymptotic formula holds. In the latter application, further specific symmetry properties of the S-matrix are derived (Lemma 7.1).

2. Definition and properties of the *S*-matrix. We consider the evolution equation

$$i\varepsilon\psi'(t) = H(t)\psi(t), \quad t \in \mathbf{R}, \quad \varepsilon \to 0, \tag{2.1}$$

where the prime denotes the derivative with respect to t, $\psi(t) \in \mathbf{C}^n$, and $H(t) \in M_n(\mathbf{C})$ for all t. We make some assumptions on the generator $H(t)$. The first is the usual analyticity condition in this context.

H1. *There exists a strip*

$$S_\alpha = \{z \in \mathbf{C} |\, |\mathrm{Im} z| \leq \alpha\}, \quad \alpha > 0, \tag{2.2}$$

such that $H(z)$ is analytic for all $z \in S_\alpha$.

Since we are studying scattering properties, we need sufficient decay at infinity.

H2. *There exist two nondegenerate matrices $H(+), H(-) \in M_n(\mathbf{C})$ and $a > 0$ such that*

$$\lim_{t \to \pm\infty} |t|^{1+a} \sup_{|s| \leq \alpha} \|H(t+is) - H(\pm)\| < \infty. \tag{2.3}$$

We finally give a condition which has to do with the physics behind the problem.

H3. *For $t \in \mathbf{R}$, the spectrum of $H(t)$, denoted by $\sigma(t)$, is real and nondegenerate*

$$\sigma(t) = \{e_1(t) < e_2(t) < \cdots < e_n(t)\} \subset \mathbf{R}, \tag{2.4}$$

and there exists $g > 0$ such that

$$\inf_{\substack{j \neq k \\ t \in \mathbf{R}}} |e_j(t) - e_k(t)| \geq g. \tag{2.5}$$

As a consequence of H3, for each $t \in \mathbf{R}$, there exists a complete set of projectors $P_j(t) = P_j^2(t) \in M_n(\mathbf{C})$, $j = 1, 2, \ldots, n$, such that

$$\sum_{j=1}^{n} P_j(t) \equiv \mathbf{I}, \tag{2.6}$$

$$H(t) = \sum_{j=1}^{n} e_j(t) P_j(t), \tag{2.7}$$

and there exists a basis of $\mathbf{C}^n$ of eigenvectors of $H(t)$. We determine these eigenvectors $\varphi_j(t)$, $j = 1, 2, \ldots, n$, uniquely (up to a constant) by requiring them to satisfy

$$H(t)\varphi_j(t) = e_j(t)\varphi_j(t), \tag{2.8}$$

$$P_j(t)\varphi_j'(t) \equiv 0, \quad j = 1, 2, \ldots, n. \tag{2.9}$$

Explicitly, if $\psi_j(t)$, $j = 1, 2, \ldots, n$, form a complete set of differentiable eigenvectors of $H(t)$, the eigenvectors

$$\varphi_j(t) = \mathrm{e}^{-\int_0^t \xi_j(t')dt'} \psi_j(t) \quad \text{s.t. } \varphi_j(0) = \psi_j(0) \tag{2.10}$$

with

$$\xi_j(t) = \frac{\langle \psi_j(t) | P_j(t) \psi_j'(t) \rangle}{\|\psi_j(t)\|^2}, \quad j = 1, \ldots, n, \tag{2.11}$$

verify (2.9). The fact that this choice leads to an analytic set of eigenvectors close to the real axis will be proven below. We expand the solution $\psi(t)$ along the basis just constructed, thus defining the unknown coefficients $c_j(t)$, $j = 1, 2, \ldots, n$, to be determined,

$$\psi(t) = \sum_{j=1}^{n} c_j(t) \mathrm{e}^{-i\int_0^t e_j(t')dt'/\varepsilon} \varphi_j(t). \tag{2.12}$$

The phases $\mathrm{e}^{-i\int_0^t e_j(t')dt'/\varepsilon}$ (see H3) are introduced for convenience. By inserting (2.12) into (2.1), we get the following differential equation for the $c_j(t)$'s:

$$c_j'(t) = \sum_{k=1}^{n} a_{jk}(t) \mathrm{e}^{i\Delta_{jk}(t)/\varepsilon} c_k(t), \tag{2.13}$$

where

$$\Delta_{jk}(t) = \int_0^t (e_j(t') - e_k(t'))dt' \tag{2.14}$$

and

$$a_{jk}(t) = -\frac{\langle \varphi_j(t) | P_j(t) \varphi_k'(t) \rangle}{\|\varphi_j(t)\|^2}. \tag{2.15}$$

Here $\langle \cdot | \cdot \rangle$ denotes the usual scalar product in $\mathbf{C}^n$. Our choice (2.9) implies $a_{jj}(t) \equiv 0$. It is also shown below that the $a_{jk}(t)$'s are analytic functions in a neighborhood of the real axis and that hypothesis H2 implies that they satisfy the estimate

$$\lim_{t\to\pm\infty} \sup_{j\neq k} |t|^{1+a} |a_{jk}(t)| < \infty. \tag{2.16}$$

As a consequence of this last property and of the fact that the eigenvalues are real by assumption, the following limits exist:

$$\lim_{t\to\pm\infty} c_j(t) = c_j(\pm\infty). \tag{2.17}$$

We are now able to define the associated S-matrix, $S \in M_n(\mathbf{C})$, by the identity

$$S \begin{pmatrix} c_1(-\infty) \\ c_2(-\infty) \\ \vdots \\ c_n(-\infty) \end{pmatrix} = \begin{pmatrix} c_1(+\infty) \\ c_2(+\infty) \\ \vdots \\ c_n(+\infty) \end{pmatrix}. \tag{2.18}$$

Such a relation makes sense because of the linearity of equation (2.13). It is a well-known result that under our general hypotheses, the S-matrix satisfies

$$S = \mathbf{I} + \mathcal{O}(\varepsilon). \tag{2.19}$$

Note that the jth column of the S-matrix is given by the solution of (2.13) at $t = \infty$ subjected to the initial conditions $c_k(-\infty) = \delta_{jk}$, $k = 1, 2, \ldots, n$.

In general, the S-matrix defined above has no particular properties besides that of being invertible. However, when the generator $H(t)$ satisfies some symmetry properties, the same is true for S. Since such properties are important in applications, we show below that if $H(t)$ is self-adjoint with respect to some indefinite scalar product, then S is unitary with respect to another indefinite scalar product. Let $J \in M_n(\mathbf{C})$ be an invertible self-adjoint matrix. We define an indefinite metric on $\mathbf{C}^n$ by means of the indefinite scalar product

$$(\cdot,\cdot)_J = \langle\cdot|J\cdot\rangle. \tag{2.20}$$

It is easy to check that the adjoint $A^{\#}$ of a matrix A with respect to the $(\cdot,\cdot)_J$ scalar product is given by

$$A^{\#} = J^{-1}A^*J. \tag{2.21}$$

Proposition 2.1. *Let $H(t)$ satisfy* H1 *and* H2 *and possess n distinct eigenvalues $\forall t \in \mathbf{R}$. Furthermore, assume that $H(t)$ is self-adjoint with respect to the scalar product $(\cdot,\cdot)_J$,*

$$H(t) = H^{\#}(t) = J^{-1}H^*(t)J, \quad \forall t \in \mathbf{R}, \tag{2.22}$$

and the eigenvectors $\varphi_j(0)$ of $H(0)$ satisfy

$$(\varphi_j(0), \varphi_j(0))_J = \rho_j, \quad \rho_j \in \{-1, 1\}, \quad \forall j = 1, \dots, n. \tag{2.23}$$

Then the eigenvalues of $H(t)$ are real $\forall t \in \mathbf{R}$ and the S-matrix is unitary with respect to the scalar product $(\cdot,\cdot)_R$, where $R = R^* = R^{-1}$ is the real diagonal matrix $R = \mathrm{diag}(\rho_1, \rho_2, \dots, \rho_n)$,

$$S^{\#} = RS^*R = S^{-1}. \tag{2.24}$$

Remark. The condition $(\varphi_j(0), \varphi_j(0))_J = \pm 1$ can always be satisfied by suitable renormalization provided $(\varphi_j(0), \varphi_j(0))_J \neq 0$.

The main interest of this proposition is that when the S-matrix possesses symmetries, some of its elements can be deduced from resulting identities without resorting to their actual computations.

A simple proof of Proposition 2.1 that makes use of notions discussed in the next section can be found in Appendix A. Proposition 2.1 can actually be used for the two main applications that we deal with in section 7. Note that in specific cases, further symmetry properties can be derived for the S-matrix; see section 7.

3. Analyticity properties. The generator $H(z)$ is analytic in S_α; hence the solution of the linear equation (2.1) $\psi(z)$ is analytic in S_α as well. However, the eigenvalues and eigenprojectors of $H(z)$ may have singularities in S_α. Let us recall some basic properties, the proofs of which can be found in [K]. The eigenvalues and eigenprojectors of a matrix analytic in a region of the complex plane have analytic continuations in that region with possible singularities located at points z_0, called exceptional points. In a neighborhood free of exceptional points, the eigenvalues are given by branches of analytic functions and their multiplicities are constant. One eigenvalue can therefore be analytically continued until it coincides at z_0 with one or several other eigenvalues. The set of such points defines the set of exceptional points. The eigenvalues may possess branching points at an exceptional z_0, where they are continuous, whereas the eigenprojectors are also multivalued but diverge as $z \to z_0$. Hence by hypothesis H3, the n distinct eigenvalues $e_j(t)$ defined on the real axis are

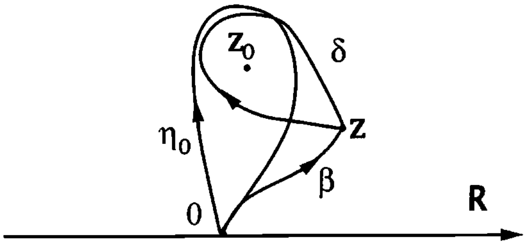


FIG. 1. *The paths β, δ, and η_0 in $S_\alpha \backslash \Omega$.*

analytic on the real axis and possess multivalued analytic continuations in S_α, with possible branching points at the set of degeneracies Ω, given by

$$\Omega = \{z_0 | \; e_j(z_0) = e_k(z_0) \quad \text{for some } k \text{ and } j \text{ and some analytic continuation}\}. \tag{3.1}$$

By assumption H2, Ω is finite, and by H3, $\Omega \cap \mathbf{R} = \emptyset$ and $\Omega = \overline{\Omega}$ due to Schwarz's principle. Similarly, the eigenprojectors $P_j(t)$ defined on the real axis are analytic on the real axis and possess multivalued analytic continuations in S_α with possible singularities at Ω. To see more precisely what happens to these multivalued functions when we turn around a point $z_0 \in \Omega$, we consider the construction described in Figure 1. Let f be a multivalued analytic function in $S_\alpha \backslash \Omega$. We denote by $f(z)$ the analytic continuation of the restriction of f around 0 along some path $\beta \in S_\alpha \backslash \Omega$ from 0 to z. Then we perform the analytic continuation of $f(z)$ along a negatively oriented loop δ based at z around a unique point $z_0 \in \Omega$, and we denote by $\widetilde{f}(z)$ the function that we get when we come back to the starting point. (If δ is positively oriented, the construction is similar.) For later purposes, we define η_0 as the negatively oriented loop homotopic to the loop based at the origin encircling z_0 obtained by following β from 0 to z, δ from z back to z, and β in the reverse sense from z back to the origin. We will keep this notation in the rest of this section. It follows from the discussion above that if we perform the analytic continuation of the set of eigenvalues $\{e_j(z)\}_{j=1}^n$, along a negatively oriented loop around $z_0 \in \Omega$, we get the set $\{\widetilde{e}_j(z)\}_{j=1}^n$ with

$$\widetilde{e}_j(z) = e_{\sigma_0(j)}(z), \quad j = 1, \dots, n, \tag{3.2}$$

where

$$\sigma_0 : \; \{1, 2, \dots, n\} \to \{1, 2, \dots, n\} \tag{3.3}$$

is a permutation that depends on η_0. Similarly, and with the same notations, we get for the analytic continuations of the projectors around z_0

$$\widetilde{P}_j(z) = P_{\sigma_0(j)}(z), \quad j = 1, \dots, n. \tag{3.4}$$

Let us consider now the eigenvectors $\varphi_j(t)$. We define $W(t)$ as the solution of

$$\begin{aligned} W'(t) &= \sum_{j=1}^n P_j'(t) P_j(t) W(t) \\ &\equiv K(t) W(t), \qquad W(0) = \mathbf{I}, \end{aligned} \tag{3.5}$$

where $t \in \mathbf{R}$. It is well known [K], [Kr] that $W(t)$ satisfies the intertwining identity

$$W(t) P_j(0) = P_j(t) W(t), \quad j = 1, 2, \dots, n, \quad \forall t \in \mathbf{R}, \tag{3.6}$$

so that if $\{\varphi_j(0)\}_{j=1}^n$ denotes a set of eigenvectors of $H(0)$, the vectors defined by

$$\varphi_j(t) = W(t)\varphi_j(0) \tag{3.7}$$

are eigenvectors of $H(t)$. Moreover, using the identity $Q(t)Q'(t)Q(t) \equiv 0$, which is true for any differentiable projector, it is easily checked that condition (2.9) is satisfied by these vectors. The generator $K(t)$ is analytic on the real axis and can be analytically continued in $S_\alpha \backslash \Omega$. Actually, $K(z)$ is single valued in $S_\alpha \backslash \Omega$. Indeed, let us consider the analytic continuation of $K(z)$ around $z_0 \in \Omega$. We get from (3.4) that

$$\widetilde{P}_j'(z) = P'_{\sigma_0(j)}(z) \tag{3.8}$$

so that

$$\begin{aligned}\widetilde{K}(z) &= \sum_{j=1}^n \widetilde{P}_j'(z)\widetilde{P}_j(z) = \sum_{j=1}^n P'_{\sigma_0(j)}(z)P_{\sigma_0(j)}(z)\\ &= \sum_{k=1}^n P_k'(z)P_k(z) = K(z).\end{aligned} \tag{3.9}$$

Consequently, $W(t)$ can be analytically continued in $S_\alpha \backslash \Omega$, where it is multivalued and satisfies both (3.5) and (3.6) with $z \in S_\alpha \backslash \Omega$ in place of $t \in \mathbf{R}$. Moreover, the relation between the analytic continuation $W(z)$ from 0 to some point $z \in S_\alpha \backslash \Omega$ and the analytic continuation $\widetilde{W}(z)$ is given by a monodromy matrix $W(\eta_0)$ such that

$$\widetilde{W}(z) = W(z)W(\eta_0), \tag{3.10}$$

where η_0 is the negatively oriented loop based at the origin which encircles only $z_0 \in \Omega$ (see Figure 1). Note also that the analytic continuation $W(z)$ is invertible in $S_\alpha \backslash \Omega$ and $W^{-1}(z)$ satisfies

$$W^{-1\prime}(z) = -W^{-1}(z)K(z), \qquad W^{-1}(0) = \mathbf{I}. \tag{3.11}$$

As a consequence, the eigenvectors (3.7) possess multivalued analytic extensions in $S_\alpha \backslash \Omega$. Indeed, it is easily seen that the analytic continuation of $\varphi_j(z)$ along a negatively oriented loop around $z_0 \in \Omega$, $\widetilde{\varphi}_j(z)$, is proportional to $\varphi_{\sigma_0(j)}(z)$. Hence we introduce the quantity $\theta_j(\eta_0) \in \mathbf{C}$ by the definition

$$\widetilde{\varphi}_j(z) = \mathrm{e}^{-i\theta_j(\eta_0)}\varphi_{\sigma_0(j)}(z), \quad j = 1, 2, \ldots, n. \tag{3.12}$$

Note that this is equivalent to $W(\eta_0)\varphi_j(0) = \mathrm{e}^{-i\theta_j(\eta_0)}\varphi_{\sigma_0(j)}(0)$ (see (3.10)). Let us consider the couplings (2.15). Using the definition (3.7), the invertibility of $W(t)$, and the identity (3.6), it is not difficult to see that we can rewrite

$$a_{jk}(t) = -\frac{\langle \varphi_j(0)|P_j(0)W(t)^{-1}K(t)W(t)\varphi_k(0)\rangle}{\|\varphi_j(0)\|^2}, \quad t \in \mathbf{R}, \tag{3.13}$$

which is analytic on the real axis and can be analytically continued in $S_\alpha \backslash \Omega$, where it is multivalued. Thus the same is true for the coefficients $c_j(t)$ which satisfy the linear differential equation (2.13), and their analytic continuations satisfy the same equation with $z \in S_\alpha \backslash \Omega$ in place of $t \in \mathbf{R}$. We now come to the main identity of this section regarding the coefficients $c_j(z)$. Let us denote by $c_j(z)$ the analytic continuation of

$c_j(0)$ from 0 to some $z \in S_\alpha \backslash \Omega$. We perform the analytic continuation of $c_j(z)$ along a negatively oriented loop around $z_0 \in \Omega$ and denote by $\widetilde{c}_j(z)$ the function that we get when we come back at the starting point z.

LEMMA 3.1. *For any $j = 1, \ldots, n$, we have*

$$\widetilde{c}_j(z) e^{-i \int_{\eta_0} e_j(u)du/\varepsilon} e^{-i\theta_j(\eta_0)} = c_{\sigma_0(j)}(z) \tag{3.14}$$

where η_0, $\theta_j(\eta_0)$ and $\sigma_0(j)$ are defined as above.

Proof. It follows from hypothesis H1 that $\psi(z)$ is analytic in S_α so that

$$\begin{aligned} &\sum_{j=1}^{n} c_j(z) \mathrm{e}^{-i \int_0^z e_j(u)du/\varepsilon} \varphi_j(z) \\ &\quad = \sum_{j=1}^{n} \widetilde{c}_j(z) \mathrm{e}^{-i \widetilde{\int_0^z e_j(u)du}/\varepsilon} \widetilde{\varphi}_j(z) \\ &\quad = \sum_{j=1}^{n} \widetilde{c}_j(z) \mathrm{e}^{-i \int_{\eta_0} e_j(u)du/\varepsilon} \mathrm{e}^{-i \int_0^z e_{\sigma_0(j)}(u)du/\varepsilon} \mathrm{e}^{-i\theta_j(\eta_0)} \varphi_{\sigma_0(j)}(z). \end{aligned} \tag{3.15}$$

We conclude by the fact that $\{\varphi_j(z)\}_{j=1}^n$ is a basis. □

Remark. It is straightforward to generalize the study of the analytic continuations around one singular point of the functions given above to the case where the analytic continuations are performed around several singular points since Ω is finite. The loop η_0 can be rewritten as a finite succession of individual loops encircling only one point of Ω so that the permutation σ_0 is given by the composition of a finite number of individual permutations. Thus the factors $\mathrm{e}^{-i\theta_j(\eta_0)}$ in (3.12) should be replaced by a product of such factors, each associated with one individual loop, and the same is true for the factors $\exp(-i \int_{\eta_0} e_j(z)dz/\varepsilon)$ in Lemma 3.1. This process is performed in the proof of Theorem 6.1.

4. Complex WKB analysis. This section is devoted to basic estimates on the coefficients $c_j(z)$ in certain domains extending to infinity in both the positive and negative directions inside the strip S_α. We first consider what happens in neighborhoods of $\pm\infty$. It follows from assumption H2 by a direct application of the Cauchy formula that (possibly by reducing α by an arbitrarily small amount)

$$\lim_{t \to \pm\infty} \sup_{|s| \leq \alpha} |t|^{1+a} \|H'(t+is)\| < \infty. \tag{4.1}$$

Hence the same is true for the single-valued matrix $K(z)$:

$$\lim_{t \to \pm\infty} \sup_{|s| \leq \alpha} |t|^{1+a} \|K(t+is)\| < \infty. \tag{4.2}$$

Let $0 < T \in \mathbf{R}$ be such that

$$\min_{z \in \Omega} \mathrm{Re} z > -T \quad \text{and} \quad \max_{z \in \Omega} \mathrm{Re} z < +T. \tag{4.3}$$

All quantities encountered so far are analytic in $S_\alpha \cap \{z||\mathrm{Re} z| > T\}$, and we denote with a "$\widetilde{\ }$" any analytic continuation in that set. As noticed earlier,

$$\widetilde{W}'(z) = K(z)\widetilde{W}(z), \quad z \in S_\alpha \cap \{z||\mathrm{Re} z| > T\} \tag{4.4}$$

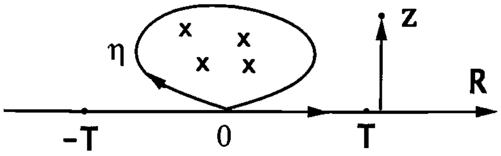


FIG. 2. *The path of integration for* $\widetilde{\Delta}_{jk}(z)$ *(the x's denote points of* Ω*).*

so that it follows from (4.2) that the limits

$$\lim_{t\to\pm\infty} \widetilde{W}(t+is) = \widetilde{W}(\pm\infty) \tag{4.5}$$

exist uniformly in $s \in]-\alpha,\alpha[$. Consequently (see (3.13)),

$$\lim_{t\to\pm\infty} |t|^{1+a} \sup_{|s|\leq\alpha} |\widetilde{a}_{jk}(t+is)| < \infty, \quad \forall j,k \in \{1,\dots,n\}. \tag{4.6}$$

Finally, for $|t| > T$, we can write

$$\begin{aligned} \mathrm{Im}\widetilde{\Delta}_{jk}(t+is) &= \mathrm{Im}\left(\int_\eta e_j(z)dz - \int_\eta e_k(z)dz\right) \\ &+ \int_0^s \mathrm{Re}(e_{\sigma(j)}(t+is') - e_{\sigma(k)}(t+is'))ds', \end{aligned} \tag{4.7}$$

where this equation is obtained by deforming the path of integration from 0 to $z = t+is$ into a loop η based at the origin, which may encircle points of Ω, followed by the real axis from 0 to Rez and a vertical path from Rez to z (see Figure 2) and σ is the corresponding permutation. Hence we have

$$\sup_{z\in S_\alpha \cap\{z||\mathrm{Re}z|>T\}} \mathrm{Im}\widetilde{\Delta}_{jk}(z) < \infty, \tag{4.8}$$

which together with (4.6) yields the existence of the limits

$$\lim_{t\to\pm\infty} \widetilde{c}_j(t+is) = \widetilde{c}_j(\pm\infty) \tag{4.9}$$

uniformly in $s \in]-\alpha,\alpha[$. We now define the domains in which useful estimates can be obtained.

DEFINITION. *Let* $j \in \{1,\dots,n\}$ *be fixed. A* dissipative domain for the index j, $D_j \subset S_\alpha \backslash \Omega$, *is such that*

$$\sup_{z\in D_j} \mathrm{Re}z = \infty, \qquad \inf_{z\in D_j} \mathrm{Re}z = -\infty \tag{4.10}$$

and is defined by the property that for any $z \in D_j$ *and any* $k \in \{1,\dots,n\}$*, there exists a path* $\gamma^k \subset D_j$ *parameterized by* $u \in]-\infty,t]$ *which links* $-\infty$ *to* z*,*

$$\lim_{u\to-\infty} \mathrm{Re}\gamma^k(u) = -\infty, \qquad \gamma^k(t) = z, \tag{4.11}$$

with

$$\sup_{z\in D_j} \sup_{u\in]-\infty,t]} \left|\frac{d}{du}\gamma^k(u)\right| < \infty, \tag{4.12}$$

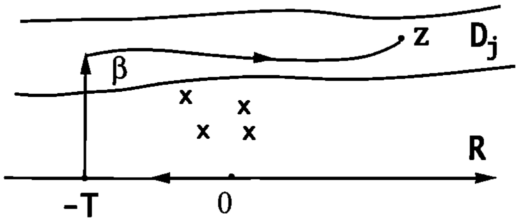


FIG. 3. *The path β along which the analytic continuation of $\Delta_{jk}(t)$ in D_j is taken.*

and satisfies the monotonicity condition

$$\text{Im}\widetilde{\Delta}_{jk}(\gamma^k(u)) \text{ is a nondecreasing function of } u \in]-\infty, t]. \tag{4.13}$$

Such a path is a dissipative path for $\{jk\}$. *Here $\widetilde{\Delta}_{jk}(z)$ is the analytic continuation of*

$$\Delta_{jk}(t) = \int_0^t (e_j(t') - e_k(t'))dt', \quad t \in \mathbf{R}, \tag{4.14}$$

in D_j along a path β described in Figure 3 going from 0 to $-T \in \mathbf{R}$ along the real axis and then vertically up or down until it reaches D_j, where $T > 0$ is chosen as in (4.3).

Let $\widetilde{c}_k(z)$, $k = 1, 2, \ldots, n$, $z \in D_j$, be the analytic continuations of $c_k(t)$ along the same path β which are solutions of the analytic continuation of (2.13) in D_j along β:

$$\widetilde{c}_k'(z) = \sum_{l=1}^{n} \widetilde{a}_{kl}(z) \mathrm{e}^{i\widetilde{\Delta}_{kl}(z)/\varepsilon} \widetilde{c}_l(z). \tag{4.15}$$

We take as initial conditions in D_j

$$\lim_{\text{Re } z \to -\infty} \widetilde{c}_k(z) = \lim_{t \to -\infty} c_k(t) = \delta_{jk}, \quad k = 1, \ldots, n, \tag{4.16}$$

and we define

$$x_k(z) = \widetilde{c}_k(z) \mathrm{e}^{i\widetilde{\Delta}_{jk}(z)/\varepsilon}, \qquad z \in D_j, \quad k = 1, \ldots, n. \tag{4.17}$$

LEMMA 4.1. *In a dissipative domain for the index j, we get the estimates*

$$\sup_{z \in D_j} |x_j(z) - 1| = \mathcal{O}(\varepsilon), \tag{4.18}$$

$$\sup_{z \in D_j} |x_k(z)| = \mathcal{O}(\varepsilon), \quad \forall k \neq j. \tag{4.19}$$

Remark. The real axis is a dissipative domain for all indices. In this case, we have $\widetilde{c}_j(t) \equiv c_j(t)$. Hence we get from the application of the lemma for all indices successively that $S = \mathbf{I} + \mathcal{O}(\varepsilon)$.

The estimates we are looking for are then just a direct corollary.

PROPOSITION 4.1. *Assume that there exists a dissipative domain D_j for the index j. Let η_j be a loop based at the origin which encircles all of the degeneracies between the real axis and D_j and let σ_j be the permutation of labels associated with η_j, in the spirit of the remark ending the previous section. The loop η_j is negatively (respectively,* *positively) oriented if D_j is above (respectively, below) the real axis. Then the solution of* (2.13) *subjected to the initial conditions $c_k(-\infty) = \delta_{jk}$ satisfies*

$$c_{\sigma_j(j)}(+\infty) = \mathrm{e}^{-i\theta_j(\eta_j)}\mathrm{e}^{-i\int_{\eta_j} e_j(z)dz/\varepsilon}\left(1 + \mathcal{O}(\varepsilon)\right), \tag{4.20}$$

$$c_{\sigma_j(k)}(+\infty) = \mathcal{O}\big(\varepsilon \mathrm{e}^{\operatorname{Im}\int_{\eta_j} e_j(z)dz/\varepsilon + h_j(e_{\sigma_j(j)}(+\infty) - e_{\sigma_j(k)}(+\infty))/\varepsilon}\big), \tag{4.21}$$

with $h_j \in [H_j^-, H_j^+]$, where $H_j^\pm$ is the maximum (respectively, minimum) imaginary part of the points at $+\infty$ in D_j:

$$H^+ = \limsup_{t\to+\infty} \sup_{s|t+is\in D_j} s, \qquad H^- = \liminf_{t\to+\infty} \inf_{s|t+is\in D_j} s. \tag{4.22}$$

Thus we see that it is possible to get the (exponentially small) asymptotic behavior of the element $s_{\sigma_j(j),j}$ of the S-matrix, provided there exists a dissipative domain for the index j. The difficult part of the problem is, of course, to prove the existence of such domains D_j, which do not necessarily exist, and to have enough of them to compute the asymptotic of the whole S-matrix. This task is the equivalent for n-level systems of studying the global behavior of the Stokes lines for two-level systems. We postpone this aspect of the problem until the next section. Note that we also get from this result an exponential bound on the elements $s_{\sigma_j(k),j}$ of the S-matrix, $k \neq j$, which may or may not be useful. If η_j encircles no point of Ω, we cannot get the asymptotic behavior of $s_{\sigma_j(j),j}$ but only get the exponential bounds. Since our main concern is asymptotic behaviors, we call the corresponding dissipative domain trivial.

Remark. In contrast with the two-level case (see [JP4]) we have to work with dissipative domains instead of working with one dissipative path for all indices. Indeed, it is not difficult to convince oneself with specific three-level cases that such a dissipative path may not exist, even when the eigenvalue degeneracies are close to the real axis. In return, we prove below the existence of dissipative domains in this situation.

Proof of Proposition 4.1. The asymptotic relation is a direct consequence of Lemma 3.1, (4.9), (4.17), and the first part of Lemma 4.1. The estimate is a consequence of the same equations, the second estimate of Lemma 4.1, and the identity, for $t > T$,

$$\begin{aligned}\operatorname{Im}\widetilde{\Delta}_{jk}(t+is) &= \operatorname{Im}\left(\int_{\eta_j} e_j(z)dz - \int_{\eta_j} e_k(z)dz\right)\\ &\quad + \int_0^s \operatorname{Re}(e_{\sigma_j(j)}(t+is') - e_{\sigma_j(k)}(t+is'))ds'.\end{aligned} \tag{4.23}$$

The path of integration from 0 to z for $\widetilde{\Delta}_{jk}(z)$ is deformed into the loop η_j followed by the real axis from 0 to Rez and a vertical path from Rez to z. It remains to take the limit $t \to +\infty$. □

Proof of Lemma 4.1. We rewrite equations (4.15) and (4.16) as an integral equation and perform an integration by parts on the exponentials:

$$\begin{aligned}\widetilde{c}_k(z) = \delta_{jk} &- i\varepsilon\sum_{l=1}^n \frac{\widetilde{a}_{kl}(z)}{\widetilde{e}_k(z) - \widetilde{e}_l(z)}\mathrm{e}^{i\widetilde{\Delta}_{kl}(z)/\varepsilon}\widetilde{c}_l(z)\\ &+ i\varepsilon\sum_{l=1}^n \int_{-\infty}^z \left(\frac{\widetilde{a}_{kl}(z')}{\widetilde{e}_k(z') - \widetilde{e}_l(z')}\right)'\mathrm{e}^{i\widetilde{\Delta}_{kl}(z')/\varepsilon}\widetilde{c}_l(z')dz'\end{aligned}$$

$$+ i\varepsilon \sum_{l,m=1}^{n} \int_{-\infty}^{z} \frac{\widetilde{a}_{kl}(z')\widetilde{a}_{lm}(z')}{\widetilde{e}_k(z') - \widetilde{e}_l(z')} \mathrm{e}^{i\widetilde{\Delta}_{km}(z')/\varepsilon} \widetilde{c}_m(z')dz'. \tag{4.24}$$

Since all eigenvalues are distinct in $S_\alpha \backslash \Omega$, the denominators are always different from 0. In terms of the functions x_k, we get

$$\begin{aligned} x_k(z) = \delta_{jk} &- i\varepsilon \sum_{l=1}^{n} \frac{\widetilde{a}_{kl}(z)}{\widetilde{e}_k(z) - \widetilde{e}_l(z)} x_l(z) \\ &+ i\varepsilon \sum_{l=1}^{n} \int_{-\infty}^{z} \left(\frac{\widetilde{a}_{kl}(z')}{\widetilde{e}_k(z') - \widetilde{e}_l(z')} \right)' \mathrm{e}^{i(\widetilde{\Delta}_{jk}(z) - \widetilde{\Delta}_{jk}(z'))/\varepsilon} x_l(z')dz' \\ &+ i\varepsilon \sum_{l,m=1}^{n} \int_{-\infty}^{z} \frac{\widetilde{a}_{kl}(z')\widetilde{a}_{lm}(z')}{\widetilde{e}_k(z') - \widetilde{e}_l(z')} \mathrm{e}^{i(\widetilde{\Delta}_{jk}(z) - \widetilde{\Delta}_{jk}(z'))/\varepsilon} x_m(z')dz'. \end{aligned} \tag{4.25}$$

We introduce the quantity

$$|||x|||_j = \sup_{\substack{z \in D_j \\ l=1,\dots,n}} |x_l(z)| \tag{4.26}$$

and consider for each k equation (4.25) along the dissipative path $\gamma^k(u)$ described in the definition of D_j such that

$$\left| \mathrm{e}^{i(\widetilde{\Delta}_{jk}(\gamma^k(t)) - \widetilde{\Delta}_{jk}(\gamma^k(u)))/\varepsilon} \right| \leq 1 \tag{4.27}$$

when $u \leq t$ along that path. Due to the integrability of the $\widetilde{a}_{kl}(z)$ at infinity and the uniform boundedness of $d\gamma^k(u)/du$, we get the estimate $|x_k(z) - \delta_{kj}| \leq \varepsilon |||x|||_j A$ for some constant A uniform in $z \in D_j$; hence $|||x|||_j \leq 1 + \varepsilon |||x|||_j A$. Consequently, for ε small enough, $|||x|||_j \leq 2$ and the result follows. □

5. Superasymptotic improvement.

All of the results above can be improved substantially by using the so-called superasymptotic renormalization method [Be], [N], [JP2]. The joint use of complex WKB analysis and superasymptotic renormalization is very powerful, as demonstrated recently in [JP4] for two-level systems, and, roughly speaking, it allows us to replace all remainders $\mathcal{O}(\varepsilon)$ by $\mathcal{O}(\mathrm{e}^{-\kappa/\varepsilon})$, where $\kappa > 0$. We briefly show how to achieve this improvement in the case of n-level systems.

Let $H(z)$ satisfy H1, H2, and H3 in S_α, and let

$$\widehat{S}_\alpha = S_\alpha \backslash \cup_{r=1,\dots,p} (J_r \cup \overline{J_r}), \tag{5.1}$$

where each J_r is an open domain containing only one point of Ω in the open upper half-plane. Hence any analytic continuation $e_j(z)$ of $e_j(t)$, $t \in \mathbf{R}$, in $\widehat{S}_\alpha$ is isolated in the spectrum of $H(z)$ so that $e_j(z)$ is analytic and multivalued in $\widehat{S}_\alpha$, and the same is true for the corresponding analytic continuation $P_j(z)$ of $P_j(t)$, $t \in \mathbf{R}$. Let σ_r be the permutation associated with the loop ζ_r based at the origin which encircles J_r once such that

$$\widetilde{e}_j(z) = e_{\sigma_r(j)}(z), \tag{5.2}$$

with the convention of section 3. The matrix $K(z)$ is analytic and single valued in $\widehat{S}_\alpha$. Consider the single-valued analytic matrix

$$H_1(z, \varepsilon) = H(z) - i\varepsilon K(z), \quad z \in \widehat{S}_\alpha. \tag{5.3}$$

For ε small enough, the spectrum of $H_1(z,\varepsilon)$ is nondegenerate $\forall z \in \widehat{S}_\alpha$ so that its eigenvalues $e_j^1(z,\varepsilon)$ and eigenprojectors $P_j^1(z,\varepsilon)$ are multivalued analytic functions in $\widehat{S}_\alpha$. Moreover, for ε small enough, the analytic continuations of $e_j^1(z,\varepsilon)$ and $P_j^1(z,\varepsilon)$ around J_r satisfy $\widetilde{e}_j^1(z) = e^1_{\sigma_r(j)}(z)$ and $\widetilde{P}_j^1(z) = P^1_{\sigma_r(j)}(z)$, as can be easily deduced from (5.2) by perturbation theory. Consequently, the matrix

$$K_1(z,\varepsilon) = \sum_{j=1}^m P_j^{1\prime}(z,\varepsilon)P_j^1(z,\varepsilon) \tag{5.4}$$

is analytic and single valued in $\widehat{S}_\alpha$. Defining the single-valued matrix

$$H_2(z,\varepsilon) = H(z) - i\varepsilon K_1(z,\varepsilon), \quad z \in \widehat{S}_\alpha, \tag{5.5}$$

we can repeat the argument for ε small enough. By induction, we set for any $q \in \mathbf{N}$

$$H_q(z,\varepsilon) = H(z) - i\varepsilon K_{q-1}(z,\varepsilon), \tag{5.6}$$

$$K_{q-1}(z,\varepsilon) = \sum_{j=1}^m P_j^{q-1\prime}(z,\varepsilon)P_j^{q-1}(z,\varepsilon), \quad z \in \widehat{S}_\alpha, \tag{5.7}$$

for ε is small enough. We have

$$H_q(z,\varepsilon) = \sum_{j=1}^m e_j^q(z,\varepsilon)P_j^q(z,\varepsilon), \tag{5.8}$$

where the eigenvalues and eigenprojections are multivalued in $\widehat{S}_\alpha$ and satisfy

$$\widetilde{e}_j^q(z,\varepsilon) = e^q_{\sigma_r(j)}(z,\varepsilon), \tag{5.9}$$

$$\widetilde{P}_j^q(z,\varepsilon) = P^q_{\sigma_r(j)}(z,\varepsilon), \quad j = 1,\dots,n, \tag{5.10}$$

with the notations of (5.2). We quote from [JP4] and [JP2] the main proposition regarding this construction.

Proposition 5.1. *Let $H(z)$ satisfy* H1, H2, *and* H3 *in S_α, and let $\widehat{S}_\alpha$ be defined as above. Then there exist constants $c > 0$ and $\varepsilon^* > 0$ and a real function $b(t)$ with $\lim_{t\to\pm\infty} |t|^{1+a}b(t) < \infty$ such that*

$$\|K_q(z,\varepsilon) - K_{q-1}(z,\varepsilon)\| \le b(\mathrm{Re}z)\varepsilon^q c^q q!, \tag{5.11}$$

$$\|K_q(z,\varepsilon)\| \le b(\mathrm{Re}z) \tag{5.12}$$

for all $z \in \widehat{S}_\alpha$, all $\varepsilon < \varepsilon^*$, and all $q \le q^*(\varepsilon) \equiv [1/\mathrm{e}c\varepsilon]$, where $[y]$ denotes the integer part of y and e *is the basis of the neperian logarithm.*

We can deduce from this that in $\widehat{S}_\alpha$

$$e_j^q(z,\varepsilon) = e_j(z) + \mathcal{O}(\varepsilon^2 b(\mathrm{Re}z)), \tag{5.13}$$

$$P_j^q(z,\varepsilon) = e_j(z) + \mathcal{O}(\varepsilon b(\mathrm{Re}z)), \quad \forall q \le q^*(\varepsilon). \tag{5.14}$$

We introduce the notation $f^{q^*(\varepsilon)} \equiv f^*$ for any quantity f^q depending on the index q, and we henceforth drop the ε in the arguments of the functions that we encounter. We define the multivalued analytic matrix $W_*(z)$ for $z \in \widehat{S}_\alpha$ by

$$W_*{}'(z) = K_*(z)W_*(z), \qquad W_*(0) = \mathbf{I}. \tag{5.15}$$

Due to the above observations and Proposition 5.1, $W_*(z)$ enjoys all of the properties that $W(z)$ does, such as

$$W_*(z)P_j^*(0) = P_j^*(z)W_*(z), \tag{5.16}$$

$$\widetilde{W}^*(z) = W_*(z)W_*(\zeta_r) \tag{5.17}$$

and, uniformly in s,

$$\lim_{t \pm \infty} W_*(t+is) = W_*(\infty). \tag{5.18}$$

Thus we define for any $z \in \widehat{S}_\alpha$ a set of eigenvectors of $H_*(z)$ by $\varphi_j^*(z) = W_*(z)\varphi_j^*(0)$, where $H_*(0)\varphi_j^*(0) = e_j^*(0)\varphi_j^*(0)$, $j = 1, \ldots, n$, that satisfy

$$\widetilde{\varphi}_j^*(0) = \exp\{-i\theta_j^*(\zeta_r)\}\varphi^*_{\sigma_r(j)}(0),$$

with $\theta_j^*(\zeta_r) = \theta(\zeta_r) + \mathcal{O}(\varepsilon) \in \mathbf{C}$. Let us expand the solution of (2.1) on this multivalued set of eigenvectors as

$$\psi(z) = \sum_{j=1}^{n} c_j^*(z) \mathrm{e}^{-i\int_0^z e_j^*(z')dz'/\varepsilon} \varphi_j^*(z). \tag{5.19}$$

Since the analyticity properties of the eigenvectors and eigenvalues of $H_*(z)$ are the same as those enjoyed by the eigenvectors and eigenvalues of $H(z)$, we get, as in Lemma 3.1,

$$\widetilde{c}_j^*(z)\mathrm{e}^{-i\int_{\zeta_r} e_j^*(u)du/\varepsilon}\mathrm{e}^{-i\theta_j^*(\zeta_r)} = c^*_{\sigma_r(j)}(z), \quad \forall z \in \widehat{S}_\alpha. \tag{5.20}$$

Substituting (5.19) in (2.1), we see that in $\widehat{S}_\alpha$ the multivalued coefficients $c_j^*(z)$ satisfy the differential equation

$$c_j^{*\prime}(z) = \sum_{k=1}^{n} a_{jk}^*(z)\mathrm{e}^{i\Delta_{jk}^*(z)/\varepsilon}c_k^*(z), \tag{5.21}$$

where

$$\Delta_{jk}^*(z) = \int_0^z e_j^*(z') - e_k^*(z')dz' \tag{5.22}$$

and

$$a_{jk}^*(z) = \frac{\langle \varphi_j^*(z)(0) | P_j^*(z)(0)W_*(z)^{-1}(K_{q^*-1}(z) - K_{q^*}(z))W_*(z)\varphi_k^*(0)\rangle}{\|\varphi_j^*(0)\|^2}; \tag{5.23}$$

compare this with (3.13). The key point of this construction is that it follows from Proposition 5.1 with $q = q^*(\varepsilon)$ that

$$|a_{jk}^*(z)| \leq 2b(\mathrm{Re}z)\mathrm{e}^{-\kappa/\varepsilon}, \quad \forall z \in \widehat{S}_\alpha, \tag{5.24}$$

where $\kappa = 1/ec > 0$, and it follows from (5.13) that

$$\mathrm{Im}\Delta_{jk}^*(z) = \mathrm{Im}\Delta_{jk}(z) + \mathcal{O}(\varepsilon^2) \tag{5.25}$$

uniformly in $z \in \widehat{S}_\alpha$. Thus we deduce from (5.24) that the limits

$$\lim_{t\to\pm\infty} c_j^*(t+is) = c_j^*(\pm\infty), \quad j = 1, \dots, n, \tag{5.26}$$

exist for any analytic continuation in $\widehat{S}_\alpha$. Moreover, along any dissipative path $\gamma^k(u)$ for $\{jk\}$, as defined above, we get from (5.25)

$$\left| \mathrm{e}^{i(\widetilde{\Delta}^*_{jk}(\gamma^k(t)) - \widetilde{\Delta}^*_{jk}(\gamma^k(u)))/\varepsilon} \right| = \mathcal{O}(1), \quad \forall u \leq t, \tag{5.27}$$

so that, reproducing the proof of Lemma 4.1, we have the following result.

LEMMA 5.1. *In a dissipative domain D_j, if $\widetilde{c}_k^*(-\infty) = c_k^*(-\infty) = \delta_{kj}$, then*

$$\widetilde{c}_j^*(z) = 1 + \mathcal{O}(e^{-\kappa/\varepsilon}), \tag{5.28}$$

$$e^{i\widetilde{\Delta}_{jk}(z)\varepsilon} \widetilde{c}_k^*(z) = \mathcal{O}(e^{-\kappa/\varepsilon}), \quad \forall k \neq j, \tag{5.29}$$

uniformly in $z \in \widehat{S}_\alpha$.

This lemma yields the following improved version of our main result.

PROPOSITION 5.2. *Under the conditions of Proposition* 4.1 *and with the same notations, if $c_k^*(-\infty) = \delta_{jk}$, then*

$$c^*_{\sigma_j(j)}(+\infty) = \mathrm{e}^{-i\theta_j^*(\eta_j)} \mathrm{e}^{-i\int_{\eta_j} e_j^*(z)dz/\varepsilon} (1 + \mathcal{O}(\mathrm{e}^{-\kappa/\varepsilon})), \tag{5.30}$$

$$c^*_{\sigma_j(k)}(+\infty) = \mathcal{O}\left(\mathrm{e}^{-\kappa/\varepsilon} \mathrm{e}^{\mathrm{Im} \int_{\eta_j} e_j(z)dz/\varepsilon + h_j(e_{\sigma_j(j)}(+\infty) - e_{\sigma_j(k)}(+\infty))/\varepsilon} \right). \tag{5.31}$$

Note that we may or may not replace $e_j(z)$ by $e_j^*(z)$ in the estimate without altering the result. It remains to make the link between the S-matrix and the $c_k^*(+\infty)$'s of the proposition explicit. We define $\beta_j^{*\pm}$ by the relations

$$\varphi_j^*(\pm\infty) = \mathrm{e}^{-i\beta_j^{*\pm}} \varphi_j(\pm\infty) \tag{5.32}$$

($H_*(z)$ and $H(z)$ coincide at $\pm\infty$). By comparison with (5.19) and (2.12), we deduce the following lemma.

LEMMA 5.2. *If $c_k(t)$ and $c_k^*(t)$ satisfy $c_k(-\infty) = c_k^*(-\infty) = \delta_{jk}$, then the element kj of the S-matrix is given by*

(5.33)

$$\begin{aligned} s_{kj} = c_k(+\infty) &= e^{-i(\beta_k^{*+} - \beta_j^{*-})} e^{-i\int_0^{+\infty} e_k^*(t') - e_k(t')dt'/\varepsilon} e^{-i\int_{-\infty}^0 e_j^*(t') - e_j(t')dt'/\varepsilon} c_k^*(+\infty) \\ &\equiv e^{-i\alpha_{kj}^*} c_k^*(+\infty), \end{aligned}$$

with $\beta_j^{*\pm} = \mathcal{O}(\varepsilon)$ and $\int_{\pm\infty}^0 e_j^*(t') - e_j(t')dt'/\varepsilon = \mathcal{O}(\varepsilon)$, i.e., $e^{-i\alpha_{kj}^*} = 1 + \mathcal{O}(\varepsilon)$.

Remarks. (i) Proposition 5.2 together with Lemma 5.2 are the main results of the first part of this paper.

(ii) As a direct consequence of these estimates on the real axis, we have

$$s_{jk} = \mathcal{O}(\mathrm{e}^{-\kappa/\varepsilon}), \quad \forall k \neq j, \tag{5.34}$$

and

$$s_{jj} = \mathrm{e}^{-i\alpha_{jj}^*}(1 + \mathcal{O}(e^{-\kappa/\varepsilon})). \tag{5.35}$$

(iii) It should be clear from the analysis performed above that all of the results obtained hold if the generator $H(z)$ in (2.1) is replaced by

$$H(z,\varepsilon) = H_0(z) + \mathcal{O}(\varepsilon b(\mathrm{Re} z)), \tag{5.36}$$

with $b(t) = \mathcal{O}(1/t^{1+a})$, provided $H_0(z)$ satisfies the hypotheses we assumed.

6. Avoided crossings. We now come to the second part of the paper, in which we prove asymptotic formulas for the off-diagonal elements of the S-matrix by means of the general setup presented above. To start with, we define a class of n-level systems for which we can prove the existence of one nontrivial dissipative domain for all indices. They are obtained by means of systems that exhibit degeneracies of eigenvalues on the real axis, hereafter called real crossings, which we perturb in such a way that these degeneracies are lifted and turn into avoided crossings on the real axis. When the perturbation is small enough, this process moves the eigenvalue degeneracies off the real axis, but they remain close to the place where the real crossings occurred. This method was used successfully in [J] to deal with two-level systems. We do not attempt to list all of the cases in which dissipative domains can be constructed by means of this technique but rather present a wide class of examples which are relevant in the theory of quantum adiabatic transitions and in the theory of multichannel semiclassical scattering, as described below.

Let $H(t,\delta) \in M_n(\mathbf{C})$ satisfy the following assumptions.

H4. *For each fixed $\delta \in [0,d]$, the matrix $H(t,\delta)$ satisfies* H1 *in a strip S_α independent of δ and $H(z,\delta)$ and $\partial/\partial z H(z,\delta)$ are continuous as a functions of two variables $(z,\delta) \in S_\alpha \times [0,d]$. Moreover, it satisfies* H2 *uniformly in $\delta \in [0,d]$, with limiting values $H(\pm,\delta)$ which are continuous functions of $\delta \in [0,d]$.*

H5. *For each $t \in \mathbf{R}$ and each $\delta \in [0,d]$, the spectrum of $H(t,\delta)$, denoted by $\sigma(t,\delta)$, consists of n real eigenvalues*

$$\sigma(t,\delta) = \{e_1(t,\delta), e_2(t,\delta), \dots, e_n(t,\delta)\} \subset \mathbf{R} \tag{6.1}$$

which are distinct when $\delta > 0$:

$$e_1(t,\delta) < e_2(t,\delta) < \cdots < e_n(t,\delta). \tag{6.2}$$

When $\delta = 0$, the functions $e_j(t,0)$ are analytic on the real axis and there exists a finite set of crossing points $\{t_1 \le t_2 \le \cdots \le t_p\} \in \mathbf{R}$, $p \ge 0$, such that the following hold:

(i) *$\forall t < t_1$,*

$$e_1(t,0) < e_2(t,0) < \cdots < e_n(t,0). \tag{6.3}$$

(ii) *$\forall j < k \in \{1,2,\dots,n\}$, there exists at most one t_r with*

$$e_j(t_r,0) - e_k(t_r,0) = 0, \tag{6.4}$$

and if such a t_r exists, we have

$$\frac{\partial}{\partial t}\left(e_j(t_r,0) - e_k(t_r,0)\right) > 0. \tag{6.5}$$

(iii) *$\forall j \in \{1,2,\dots,n\}$, the eigenvalue $e_j(t,0)$ crosses eigenvalues whose indices are all superior to j or all inferior to j.*

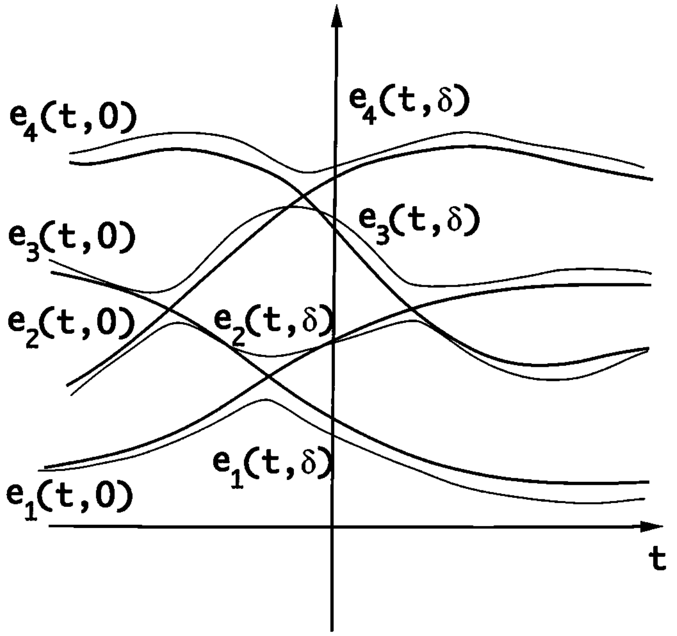


FIG. 4. *A pattern of eigenvalue crossings (bold curves) with the corresponding pattern of avoided crossings (fine curves) satisfying* H5.

Remarks. (i) The parameter δ can be understood as a coupling constant that controls the strength of the perturbation.

(ii) The eigenvalues $e_j(t,0)$ are assumed to be analytic on the real axis, because of the degeneracies on the real axis. However, if $H(t,\delta)$ is self-adjoint for any $\delta \in [0,d]$, this is true for an indexation, as follows from a theorem of Rellich; see [K].

(iii) In Figure 4, we give an example of a pattern of crossings with the corresponding pattern of avoided crossings for which the above conditions are fulfilled.

(iv) The crossings are assumed to be generic in the sense that the derivatives of $e_j - e_k$ are nonzero at the crossing t_r.

(v) The crossing points $\{t_1, t_2, \dots, t_p\}$ need not be distinct, which is important when the eigenvalues possess symmetries. However, for each $j = 1, \dots, n$, the eigenvalue $e_j(t,\delta)$ experiences avoided crossings with $e_{j+1}(t,\delta)$ and/or $e_{j-1}(t,\delta)$ at a subset of distinct points $\{t_{r_1}, \dots, t_{r_j}\} \subseteq \{t_1, t_2, \dots, t_p\}$.

We now state the main lemma of this section regarding the analyticity properties of the perturbed levels and the existence of dissipative domains for all indices in this perturbative context.

LEMMA 6.1. *Let* $H(t,\delta)$ *satisfy* H4 *and* H5. *We can choose* $\alpha > 0$ *small enough so that the following assertions are true for sufficiently small* $\delta > 0$*:*

(i) *Let* $\{t_{r_1}, \dots, t_{r_j}\}$ *be the set of avoided crossing points experienced by* $e_j(t,\delta)$, $j = 1, \dots, n$. *For each* j, *there exists a set of distinct domains* $J_r \in S_\alpha$, *where* $r \in \{r_1, \dots, r_j\}$,

$$J_r = \{z = t + is \,|\, 0 \le |t - t_r| < L, \quad 0 < g < s < \alpha'\}, \tag{6.6}$$

with L *small enough,* $\alpha' < \alpha$, *and* $g > 0$ *such that* $e_j(-\infty,\delta)$ *can be analytically* *continued in*

$$S_\alpha^j = S_\alpha \setminus \cup_{r=r_1,\ldots,r_j} (J_r \cup \overline{J_r}). \tag{6.7}$$

(ii) *Let t_r be an avoided crossing point of $e_j(t,\delta)$ with $e_k(t,\delta)$, $k = j \pm 1$. Then the analytic continuation of the restriction of $e_j(t,\delta)$ around t_r along a loop based at $t_r \in \mathbf{R}$ which encircles J_r once yields $\widetilde{e}_j(t_r,\delta)$ back at t_r with*

$$\widetilde{e}_j(t_r,\delta) = e_k(t_r,\delta). \tag{6.8}$$

(iii) *For each $j = 1,\ldots,n$, there exists a dissipative domain D_j above or below the real axis in $S_\alpha \cap \{z = t + is| \, |s| \geq \alpha'\}$. The permutation σ_j associated with these dissipative domains* (*see Proposition* 4.1) *are all given by $\sigma_j = \sigma$, where σ is the permutation that maps the index of the kth eigenvalue $e_j(\infty,0)$ numbered from the lowest one on k for all $k \in \{1,2,\cdots,n\}$.*

Remarks. (i) In part (ii), the same result is true along a loop encircling $\overline{J_r}$.

(ii) The dissipative domains D_j of part (iii) are located above (respectively, below) all of the sets J_r (respectively, $\overline{J_r}$), $r = 1,\ldots,p$.

(iii) The main interest of this lemma is that the sufficient conditions required for the existence of dissipative domains in the complex plane can be deduced from the behavior of the eigenvalues on the *real* axis.

(iv) We emphasize that more general types of avoided crossings than those described in H5 may lead to the existence of dissipative domains for *certain* indices, but we want to obtain dissipative domains for *all* indices. For example, if part (iii) of H5 is satisfied for certain indices only, then part (iii) of Lemma 6.1 is satisfied for those indices only.

(v) Note also that there are patterns of eigenvalue crossings for which there exist no dissipative domain for some indices. For example, if $e_j(t,0)$ and $e_k(t,0)$ display two crossings, it is not difficult to see from the proof of the lemma that no dissipative domains can exist for j or k.

We postpone the proof of Lemma 6.1 to the end of this section and continue with its consequences. By applying the results of the previous section, we get the following result.

THEOREM 6.1. *Let $H(t,\delta)$ satisfy* H4 *and* H5. *If $\delta > 0$ is small enough, the elements $\sigma(j)j$ of the S-matrix, with $\sigma(j)$ defined in Lemma* 6.1, *are given in the limit $\varepsilon \to 0$ for all $j = 1,\ldots,n$ by*

$$s_{\sigma(j)j} = \prod_{k=j}^{\sigma(j)\mp 1} e^{-i\theta_k(\zeta_k)} e^{-i\int_{\zeta_k} e_k(z,\delta)dz/\varepsilon} \left(1 + \mathcal{O}(\varepsilon)\right), \quad \sigma(j) \begin{cases} > j, \\ < j, \end{cases} \tag{6.9}$$

where for $\sigma(j) > j$ (respectively, $\sigma(j) < j$), ζ_k, $k = j,\ldots,\sigma(j)-1$ (respectively, $k = j,\ldots,\sigma(j)+1$), denotes a negatively (respectively, positively) oriented loop based at the origin which encircles the set J_r (respectively, $\overline{J_r}$) corresponding to the avoided crossing between $e_k(t,\delta)$ and $e_{k+1}(t,\delta)$ (respectively, $e_{k-1}(z,\delta)$) at t_r, $\int_{\zeta_k} e_k(z,\delta)dz$ denotes the integral along ζ_k of the analytic continuation of $e_k(0,\delta)$, and $\theta_k(\zeta_k)$ is the corresponding factor defined by (3.12); *see Figure* 5.

More accurately, with the notations of section 5, *we have the improved formula*

$$s_{\sigma(j)j} = e^{-i\alpha^*_{\sigma(j)j}} \prod_{k=j}^{\sigma(j)\mp 1} e^{-i\theta_k^*(\zeta_k)} e^{-i\int_{\zeta_k} e_k^*(z,\delta)dz/\varepsilon} (1 + \mathcal{O}(e^{-\kappa/\varepsilon})), \quad \sigma(j) \begin{cases} > j, \\ < j. \end{cases} \tag{6.10}$$

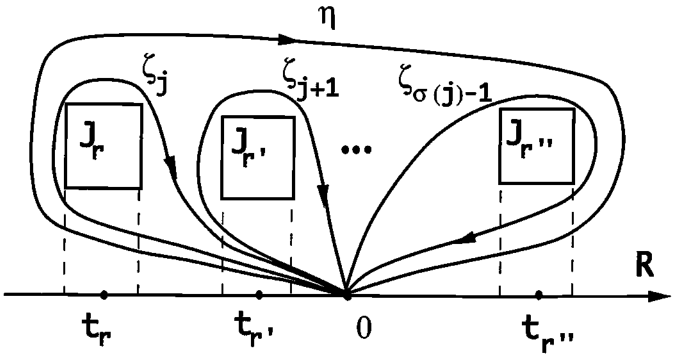


FIG. 5. *The loops η_j and ζ_k, $k = j, \dots, \sigma(j)-1$.*

The elements $\sigma(l)j$, $l \neq j$, are estimated by

$$s_{\sigma(l)j} = \mathcal{O}\left(\varepsilon e^{h(e_{\sigma(j)}(\infty,\delta)-e_{\sigma(l)}(\infty,\delta))/\varepsilon} \prod_{k=j}^{\sigma(j)\mp 1} e^{\mathrm{Im}\int_{\zeta_k} e_k(z,\delta)dz/\varepsilon}\right), \quad \sigma(j)\begin{cases} > j, \\ < j, \end{cases} \tag{6.11}$$

where h is strictly positive (*respectively, negative*) *for $\sigma(j) > j$* (*respectively, $\sigma(j) < j$*)*.*

Remarks. (i) Since the eigenvalues are continuous at the degeneracy points, we have that

$$\lim_{\delta\to 0} \mathrm{Im} \int_{\zeta_k} e_k(z,\delta)dz = 0, \quad \forall k = 1, \dots, p. \tag{6.12}$$

(ii) The remainders $\mathcal{O}(\varepsilon)$ depend on δ, but it should be possible to get estimates that are valid as both ε and δ tend to zero, in the spirit of [J], [MN], and [R].

(iii) This result shows that at least one off-diagonal element per column of the S-matrix can be computed asymptotically. However, it is often possible to get more elements by making use of the symmetries of the S-matrix. Moreover, if there exist dissipative domains that go above or below other eigenvalue degeneracies further away in the complex plane, other elements of the S-matrix can be computed.

(iv) Finally, note that all starred quantities in (6.10) depend on ε.

Proof of Theorem 6.1. The first thing to determine is whether the loops ζ_k are above or below the real axis. Since the formulas that we deduce from the complex WKB analysis are asymptotic, it suffices to choose the case that yields exponential decay of $s_{\sigma(j)j}$. It is readily checked in the proof of Lemma 6.1 below that if $\sigma(j) > j$, D_j is above the real axis and if $\sigma(j) < j$, D_j is below the real axis. Then it remains to explain how to pass from the loop η_j given in Proposition 4.1 to the set of loops ζ_k, $k = j, \dots, \sigma(j)-1$. We briefly deal with the case where $\sigma(j) > j$; the other case is similar. It follows from Lemma 6.1 that we can deform η_j into the set of loops ζ_k, each associated with one avoided crossing, as described in Figure 5. Thus we have

$$\int_{\eta_j} = \sum_{k=j}^{\sigma(j)-1} \int_{\zeta_k} \tag{6.13}$$

for the decay rate and (see (3.10))

$$W(\eta_j) = W(\zeta_{\sigma(j)-1}) \cdots W(\zeta_{j+1}) W(\zeta_j) \tag{6.14}$$

for the prefactors. Let ν_j be a negatively oriented loop based at t_r which encircles J_r as described in Lemma 6.1. Now consider the loop ζ_j associated with this avoided crossing and deform it to the path obtained by going from 0 to t_r along the real axis, from t_r to t_r along ν_j, and back from t_r to the origin along the real axis. By point (ii) of Lemma 6.1, we get

$$\widetilde{e}_j(0,\delta) = e_{j+1}(0,\delta) \tag{6.15}$$

along ζ_j, and, accordingly (see (3.12)),

$$\widetilde{\varphi}_j(0,\delta) = \mathrm{e}^{-i\theta_j(\zeta_j)}\varphi_{j+1}(0,\delta). \tag{6.16}$$

This justifies the first factor in the formula. By repeating the argument at the next avoided crossings, keeping in mind that we get $e_{j+1}(0,\delta)$ at the end of ζ_j and so on, we get the final result. The estimate on $s_{\sigma(l)j}$ is obtained by direct application of lemma 6.1. □

Proof of Lemma 6.1. In what follows, we shall denote "$\frac{\partial}{\partial t}$" by a "$\prime$." We must consider the analyticity properties of $\widetilde{e}_j(z,\delta)$ and define domains in which every point z can be reached from $-\infty$ by means of a path $\gamma(u)$, $u \in\,]-\infty,t]$, $\gamma(t) = z$ such that $\mathrm{Im}\widetilde{\Delta}_{jk}(\gamma(u),\delta)$ is nondecreasing in u for certain indices $j \neq k$ when $\delta > 0$ is fixed. Note that by Schwarz's principle, if $\gamma(u)$ is dissipative for $\{jk\}$, then $\overline{\gamma(u)}$ is dissipative for $\{kj\}$. When $\gamma(u) = \gamma_1(u) + i\gamma_2(u)$ is differentiable, saying that $\gamma(u)$ is dissipative for $\{jk\}$ is equivalent to

$$\begin{aligned}\mathrm{Re}(\widetilde{e}_j(\gamma(u),\delta) - \widetilde{e}_k(\gamma(u),\delta))\dot{\gamma}_2(u) + \mathrm{Im}(\widetilde{e}_j(\gamma(u),\delta) - \widetilde{e}_k(\gamma(u),\delta))\dot{\gamma}_1(u) \geq 0,\\ \forall u \in\,]-\infty,t],\end{aligned} \tag{6.17}$$

where "$\dot{}$" denotes the derivative with respect to u. Moreover, if the eigenvalues are analytic in a neighborhood of the real axis, we have in that neighborhood the relation

$$\mathrm{Im}(\widetilde{e}_j(t+is,\delta) - \widetilde{e}_k(t+is,\delta)) = \int_0^s \mathrm{Re}\ (\widetilde{e}'_j(t+is',\delta) - \widetilde{e}'_k(t+is',\delta))ds', \tag{6.18}$$

which is a consequence of the Cauchy–Riemann identity. We proceed as follows. We construct dissipative domains above and below the real axis when $\delta = 0$, and we show that they remain dissipative for the perturbed quantities $\widetilde{\Delta}_{jk}(z,\delta)$, provided δ is small enough. We introduce some quantities to be used in the construction. Let $C_r \subset \{1,\dots,n\}^2$ denote the set of distinct couples of indices such that the corresponding eigenvalues experience one crossing at $t = t_r$. Similarly, $N \subset \{1,\dots,n\}^2$ denotes the set of couples of indices such that the corresponding eigenvalues never cross.

Let $I_r = [t_r - L, t_r + L] \in \mathbf{R}$, $r = 1,\dots,p$, with L so small that

$$\min_{r\in\{1,\dots,p\}}\ \min_{\{jk\}\in C_r,\, j<k}\ \inf_{t\in I_r}(e'_j(t,0) - e'_k(t,0)) \equiv 4c > 0. \tag{6.19}$$

This relation defines the constant c, and we also define b by

$$\min_{r\in\{1,\dots,p\}}\ \min_{\{jk\}\in C_r,\, j<k}\ \inf_{t\in\mathbf{R}\setminus I_r}|e_j(t,0) - e_k(t,0)| \geq 4b > 0, \tag{6.20}$$

$$\min_{\{jk\}\in N,\, j<k}\ \inf_{t\in\mathbf{R}}|e_j(t,0) - e_k(t,0)| \geq 4b > 0. \tag{6.21}$$

We further introduce

$$I_r^\alpha = \{z = t + is | t \in I_r, |s| \leq \alpha\}, \quad r = 1,\dots,p. \tag{6.22}$$

Then we choose α small enough so that the only points of degeneracy of eigenvalues in S_α are on the real axis and

$$\min_{r\in\{1,\dots,p\}}\ \min_{\{jk\}\in C_r,\, j<k}\ \inf_{z\in I_r^\alpha} \mathrm{Re}(e_j'(z,0)-e_k'(z,0)) > 2c > 0 \tag{6.23}$$

$$\min_{r\in\{1,\dots,p\}}\ \min_{\{jk\}\in C_r,\, j<k}\ \inf_{z\in S_\alpha\setminus I_r^\alpha} |\mathrm{Re}(e_j(z,0)-e_k(z,0))| > 2b > 0 \tag{6.24}$$

$$\min_{\{jk\}\in N,\, j<k}\ \inf_{z\in S_\alpha} |\mathrm{Re}(e_j(z,0)-e_k(z,0))| > 2b > 0. \tag{6.25}$$

The fact that this choice is always possible is a consequence of the analyticity of $e_j(z,0)$ close to the real axis and of the fact that we can essentially work in a compact because of hypothesis H4. Let $a(t)$ be integrable on $\mathbf{R}$ and such that

$$\frac{a(t)}{2} > \max_{j<k\in\{1,\dots,n\}}\ \sup_{|s|\le\alpha} \left|\mathrm{Re}(e_j'(t+is,0)-e_k'(t+is,0))\right|. \tag{6.26}$$

It follows from H4 that such functions exist.

Let $r\in\{1,\dots,p\}$ and $\gamma_2(u)$ be a solution of

$$\begin{cases} \dot\gamma_2(u) = -\frac{\gamma_2(u)a(u)}{b}, & u\in\,]-\infty,t_r-L],\\ \dot\gamma_2(u)=0, & u\in\,]t_r-L,t_r+L[,\\ \dot\gamma_2(u) = +\frac{\gamma_2(u)a(u)}{b}, & u\in[t_r+L,\infty[, \end{cases} \tag{6.27}$$

with $\gamma_2(t_r)>0$. Then $\gamma_2(u)>0$ for any u since

$$\begin{cases} \gamma_2(u) = \gamma_2(t_r)\mathrm{e}^{-\int_{t_r-L}^{u} a(u')du'/b}, & u\in\,]-\infty,t_r-L],\\ \gamma_2(u) = \gamma_2(t_r), & u\in\,]t_r-L,t_r+L[,\\ \gamma_2(u) = \gamma_2(t_r)\mathrm{e}^{\int_{t_r+L}^{u} a(u')du'/b}, & u\in[t_r+L,\infty[, \end{cases} \tag{6.28}$$

and since $a(u)$ is integrable, the limits

$$\lim_{u\to\pm\infty}\gamma_2(u) = \gamma_2(\pm\infty) \tag{6.29}$$

exist. Moreover, we can always choose $\gamma_2(t_r)>0$ sufficiently small so that $\gamma^r(u)\equiv u+i\gamma_2(u)\in S_\alpha$ for any real u. Let us verify that this path is dissipative for all $\{jk\}\in C_r$, $j<k$. For $u\in\,]-\infty,t_r-L]$, using

$$\mathrm{Re}(e_j(z,0)-e_k(z,0)) < -2b < 0, \quad \forall z\in S_\alpha\cap\{z|\mathrm{Re}z\le t_r-L\}, \tag{6.30}$$

$$|\mathrm{Im}(e_j(t+is,0)-e_k(t+is,0))| < |s| \sup_{s'\in[0,s]} \left|\mathrm{Re}(e_j'(t+is',0)-e_k'(t+is',0))\right| \tag{6.31}$$

(see (6.18)), and the definition (6.26), we have

$$\begin{aligned} &\mathrm{Re}(e_j(\gamma^r(u),0)-e_k(\gamma^r(u),0))\dot\gamma_2(u) + \mathrm{Im}(e_j(\gamma^r(u),0)-e_k(\gamma^r(u),0))\dot\gamma_1(u)\\ &\quad = -\mathrm{Re}(e_j(\gamma^r(u),0)-e_k(\gamma^r(u),0))\frac{\gamma_2(u)a(u)}{b} + \mathrm{Im}(e_j(\gamma^r(u),0)-e_k(\gamma^r(u),0))\\ &\quad > 2\gamma_2(u)a(u) - \gamma_2(u)a(u)/2 > \gamma_2(u)a(u) > 0. \end{aligned} \tag{6.32}$$

Similarly, when $u \geq t_r + L$, using

$$\operatorname{Re}(e_j(z,0) - e_k(z,0)) > 2b > 0, \quad \forall z \in S_\alpha \cap \{z | \operatorname{Re} z \geq t_r + L\}, \tag{6.33}$$

we get

$$\begin{aligned} &\operatorname{Re}(e_j(\gamma^r(u),0) - e_k(\gamma^r(u),0))\dot{\gamma}_2(u) + \operatorname{Im}(e_1(\gamma^r(u),0) - e_k(\gamma^r(u),0))\dot{\gamma}_1(u) \\ &\quad = \operatorname{Re}(e_j(\gamma^r(u),0) - e_k(\gamma^r(u),0))\frac{\gamma_2(u)a(u)}{b} + \operatorname{Im}(e_j(\gamma^r(u),0) - e_k(\gamma^r(u),0)) \\ &\quad > 2\gamma_2(u)a(u) - \gamma_2(u)a(u)/2 > \gamma_2(u)a(u) > 0. \end{aligned} \tag{6.34}$$

Finally, for $s \in [t_r - L, t_r + L]$, we have with (6.23) that

$$\begin{aligned} \operatorname{Im}(e_j(\gamma^r(u),0) - e_k(\gamma^r(u),0)) &= \int_0^{\gamma_2(u)} \operatorname{Re}(e'_j(t'+is,0) - e'_k(t'+is,0)) \\ &\geq \gamma_2(u)2c > \gamma_2(u)c > 0. \end{aligned} \tag{6.35}$$

Thus $\gamma^r(u)$ is dissipative for all $\{jk\} \in C_r$, $j < k$. Note that the last estimate shows that it is not possible to find a dissipative path for $\{jk\} \in C_r$, $j < k$ below the real axis.

Now consider $\{jk\} \in N$, $j < k$, and let $\gamma_2^+(u)$ be a solution of

$$\dot{\gamma}_2^+(u) = -\frac{\gamma_2^+(u)a(u)}{b}, \qquad \gamma_2^+(0) > 0, \quad u \in]-\infty, +\infty[, \tag{6.36}$$

i.e.,

$$\gamma_2^+(u) = \gamma_2^+(0)\mathrm{e}^{-\int_0^u a(u')du'/b}. \tag{6.37}$$

As above, we have $\gamma_2^+(u) > 0$ for any u and we can choose $\gamma_2^+(0) > 0$ small enough so that $\gamma^+(u) \equiv u + i\gamma_2^+(u) \in S_\alpha$ for any $u \in \mathbf{R}$. Since

$$\operatorname{Re}(e_j(z,0) - e_k(z,0)) > -2b, \quad \forall z \in S_\alpha, \tag{6.38}$$

we check by a computation analogous to (6.32) that $\gamma^+(u)$ is dissipative for $\{jk\} \in N$, $j < k$. Similarly, we can verify that if $\gamma_2^-(u)$ is the solution of

$$\dot{\gamma}_2^-(u) = \frac{\gamma_2^-(u)a(u)}{b}, \qquad \gamma_2^-(0) < 0, \quad u \in]-\infty, +\infty[, \tag{6.39}$$

with $|\gamma_2^-(0)|$ small enough, the path $\gamma^-(u) \equiv u + i\gamma_2^-(u)$ below the real axis is in S_α for any $u \in \mathbf{R}$ and is dissipative for $\{jk\} \in N$, $j < k$, as well.

Finally, the complex conjugates of these paths yield dissipative paths above and below the real axis for $\{jk\} \in N$, $j > k$.

We now define the dissipative domains by means of their borders. Let $\gamma^+(u)$ and $\gamma^-(u)$, $u \in \mathbf{R}$, be two dissipative paths in S_α defined as above with $|\gamma_2^-(0)|$ sufficiently small so that $\overline{\gamma^-}$ is below γ^+. We set

$$D = \{z = t + is | 0 < -\gamma_2^-(t) \leq s \leq \gamma_2^+(t), \quad t \in \mathbf{R}\}. \tag{6.40}$$

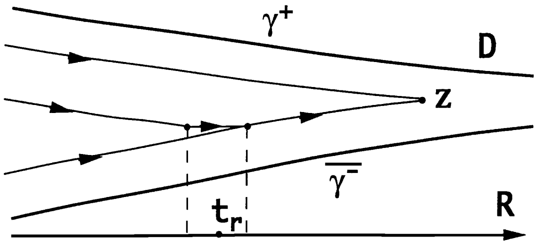


FIG. 6. *The dissipative domain D and some dissipative paths.*

Let $z \in D$, and $j \in \{1, \dots, n\}$ be fixed. By assumption H5, the set X_j of indices k such that $\{jk\} \in C_r$ for some $r \in \{1, \dots, p\}$ consists of values k that satisfy $j < k$ or it consists of values k that satisfy $j > k$. Let us assume that the first alternative takes place. Now for any $k \in \{1, \dots, n\}$, there are three cases.

(1) If $k \in X_j$, then there exists a dissipative path $\gamma^r \in D$ for $\{jk\} \in C_r$, $j < k$, constructed as above which links $-\infty$ to z. It is enough to select the initial condition $\gamma_2(t_r)$ suitably; see Figure 6.

(2) Similarly, if $j < k \notin X_j$, there exists a dissipative path $\gamma^+ \in D$ for $\{jk\}$ constructed as above which links $-\infty$ to z obtained by a suitable choice of $\gamma_2^+(0)$.

(3) Finally, if $k > j$, we can take as a dissipative path for $\{jk\}$ the path $\overline{\gamma^-} \in D$ constructed as above which links $-\infty$ to z with a suitable choice of $\gamma_2^-(0)$. Hence D is dissipative for the index j when $\delta = 0$. If j is such that the set X_j consists of points k with $k > j$, a similar argument with the complex conjugates of the above paths shows that the domain $\overline{D}$ below the real axis is dissipative for j when $\delta = 0$.

Let us show that these domains remain dissipative when $\delta > 0$ is not too large. We start by considering the analyticity properties of the perturbed eigenvalues $e_j(z, \delta)$, $\delta > 0$. Let $0 < \alpha' < \alpha$ be such that

$$I_r^{\alpha'} \cap (D \cup \overline{D}) = \emptyset, \quad \forall r = 1, \dots, p. \tag{6.41}$$

The analytic eigenvalues $e_j(z, 0)$, $j \in \{1, \dots, n\}$, are isolated in the spectrum of $H(z, 0)$ for any $z \in \widetilde{S}_\alpha$, where

$$\widetilde{S}_\alpha = S_\alpha \setminus \cup_{r=1,\dots,p} I_r^{\alpha'}. \tag{6.42}$$

For any $j = 1, \dots, n$ we get from perturbation theory [K] that the analytic continuations $\widetilde{e}_j(z, \delta)$ of $e_j(t_1 - L, \delta)$ in $\widetilde{S}_\alpha$ are all distinct in $\widetilde{S}_\alpha$, provided δ is small enough. This is due to the fact that assumption H4 implies the continuity of $H(z, \delta)$ in δ uniformly in $z \in S_\alpha$, as is easily verified. More precisely, for any fixed index j, the eigenvalue $e_j(t, \delta)$ experiences avoided crossings at the points $\{t_{r_1}, \dots, t_{r_j}\}$. We can assume without loss of generality that

$$I_k^{\alpha'} \cap I_l^{\alpha'} = \emptyset, \quad \forall k \neq l \in \{r_1, \dots, r_j\}. \tag{6.43}$$

Hence for $\delta > 0$ small enough, the analytic continuation $\widetilde{e}_j(z, \delta)$ is isolated in the spectrum of $H(z, \delta)$ uniformly in $z \in S_\alpha \setminus \cup_{r=r_1,\dots,r_j} I_r^{\alpha'}$. Since by assumption H5 there is no crossing of eigenvalues on the real axis when $\delta > 0$, there exists a $0 < g < \alpha'$ that depends on δ such that $\widetilde{e}_j(z, \delta)$ is isolated in the spectrum of $H(z, \delta)$ uniformly in $z \in S_\alpha^j$, where

$$S_\alpha^j = S_\alpha \setminus \cup_{r=r_1,\dots,r_j} (J_r \cup \overline{J_r}) \tag{6.44}$$

and

$$J_r = I_r^{\alpha'} \cap \{z|\,\mathrm{Im}z > g\}, \quad r = 1, \dots, p. \tag{6.45}$$

Hence the singularities of $\widetilde{e}_j(z,\delta)$ are located in $\cup_{r=r_1,\dots,r_j}(J_r \cup \overline{J_r})$, which yields the first assertion of the lemma.

Consider a path ν_r from $t_r - L$ to $t_r + L$ which goes above J_r, where t_r is an avoided crossing between $e_j(t,\delta)$ and $e_k(t,\delta)$, $k = j \pm 1$. By perturbation theory again, $e_j(t_r - L,\delta)$ and $e_k(t_r - L,\delta)$ tend to $e_{j'}(t_r - L, 0)$ and $e_{k'}(t_r - L, 0)$ as $\delta \to 0$ for some $j', k' \in 1, \dots, n$, whereas $e_j(t_r + L,\delta)$ and $e_k(t_r + L,\delta)$ tend to $e_{k'}(t_r + L, 0)$ and $e_{j'}(t_r + L, 0)$ as $\delta \to 0$; see Figure 4. Now the analytic continuations of the restrictions of $e_j(t,\delta)$ and $e_k(t,\delta)$ around $t_r - L$ along ν_r, $\widetilde{e}_j(z,\delta)$ and $\widetilde{e}_k(z,\delta)$ tend to the analytic functions $\widetilde{e}_{j'}(z,0) = e_{j'}(z,0)$ and $\widetilde{e}_{k'}(z,0) = e_{k'}(z,0)$ as $\delta \to 0$ for all $z \in \nu_r$. Thus we deduce that for δ small enough,

$$\widetilde{e}_j(t_r + L,\delta) \equiv e_k(t_r + L,\delta) \tag{6.46}$$

since we know that $\widetilde{e}_j(t_r + L,\delta) = e_{\sigma(j)}(t_r + L,\delta)$ for some permutation σ. Hence point (iii) of the lemma follows.

Note that the analytic continuations $\widetilde{e}_j(z,\delta)$ are single valued in $\widetilde{S}_\alpha$. Indeed, the analytic continuation of $e_j(t_r - L,\delta)$ along $\overline{\nu_r}$, denoted by $\widehat{e}_j(z,\delta)$, $\forall z \in \overline{\nu_r}$, is such that

$$\widehat{e}_j(t_r + L,\delta) = \overline{\widetilde{e}_j(t_r + L,\delta)} = \widetilde{e}_j(t_r + L,\delta) = e_k(t_r + L,\delta) \tag{6.47}$$

due to Schwarz's principle. We further require δ to be sufficiently small so that the following estimates are satisfied:

$$\min_{r\in\{1,\dots,p\}} \min_{\substack{\{jk\}\in C_r\\ j<k}} \inf_{z\in\widetilde{S}_\alpha\setminus I_r^\alpha} |\mathrm{Re}(\widetilde{e}_j(z,\delta) - \widetilde{e}_k(z,\delta))| > b > 0, \tag{6.48}$$

$$\min_{\substack{\{jk\}\in N\\ j<k}} \inf_{z\in\widetilde{S}_\alpha} |\mathrm{Re}(\widetilde{e}_j(z,\delta) - \widetilde{e}_k(z,\delta))| > b > 0, \tag{6.49}$$

$$\max_{j<k\in\{1,\dots,n\}} \sup_{\mathrm{Im}z|\; z\in\widetilde{S}_\alpha} \left|\mathrm{Re}(\widetilde{e}'_j(z,\delta) - \widetilde{e}'_k(z,\delta))\right| < a(\mathrm{Re}z), \tag{6.50}$$

and, in the compacts $\widetilde{I}_r^\alpha = I_r^\alpha \setminus I_r^{\alpha'}$,

$$\begin{aligned} &\min_{r\in\{1,\dots,p\}} \min_{\substack{\{jk\}\in C_r\\ j<k}} \inf_{z\in\widetilde{I}_r^\alpha} |\mathrm{Im}(\widetilde{e}_j(z,\delta) - \widetilde{e}_k(z,\delta))| \\ &> \frac{1}{2} \min_{r\in\{1,\dots,p\}} \min_{\substack{\{jk\}\in C_r\\ j<k}} \inf_{z\in\widetilde{I}_r^\alpha} |\mathrm{Im}(\widetilde{e}_j(z,0) - \widetilde{e}_k(z,0))| > |\mathrm{Im}z|c, \end{aligned} \tag{6.51}$$

$$\begin{aligned} &\max_{r\in\{1,\dots,p\}} \max_{j<k\in\{1,\dots,n\}} \sup_{z\in\widetilde{I}_r^\alpha} |\mathrm{Im}(\widetilde{e}_j(z,\delta) - \widetilde{e}_k(z,\delta))| \\ &< 2 \max_{r\in\{1,\dots,p\}} \max_{j<k\in\{1,\dots,n\}} \sup_{z\in\widetilde{I}_r^\alpha} |\mathrm{Im}(\widetilde{e}_j(z,0) - \widetilde{e}_k(z,0))| < |\mathrm{Im}z|a(\mathrm{Re}z). \end{aligned} \tag{6.52}$$

The simultaneous requirements (6.26) and (6.50) are made possible by the continuity properties of $H'(z,\delta)$ and the uniformity in δ of the decay at $\pm\infty$ of $H(z,\delta)$ assumed in H4 together with the fact that $a(t)$ can be replaced by a multiple of $a(t)$ if necessary to satisfy both estimates. The condition on δ is given by the first inequalities in (6.51) and (6.52), whereas the second ones are just recalls.

Then it remains to check that the paths γ^r, γ^+, and γ^- defined above satisfy the dissipativity condition (6.17) for the corresponding indices. This is not difficult since the above estimates are precisely designed to preserve inequalities such as (6.32), (6.34), and (6.35). However, it should not be forgotten that in the sets $I_r^{\alpha'}$, the eigenvalues may be singular so that (6.18) cannot be used there. Therefore, when checking that a path parameterized as above by $u \in \mathbf{R}$ is dissipative, it is necessary to consider separately the case $u \in \mathbf{R}\backslash(\cup_{r=1,\ldots,p} I_r)$, where we proceed as above with (6.48), (6.49), (6.50), and (6.18), and the case $u \in \cup_{r=1,\ldots,p} I_r$, where we use use (6.51) and (6.52) instead of (6.18) as follows. If $u \in I_r$ for r such that t_r is a crossing point for $e_j(t,0)$ and $e_k(t,0)$, we take (6.51) to estimate $\text{Im}(\widetilde{e}_{j'}(z,\delta) - \widetilde{e}_{k'}(z,\delta))$ for the corresponding indices j' and k', and if t_r is not a crossing point for $e_j(t,0)$ and $e_k(t,0)$, we use (6.52) to estimate $\text{Im}(\widetilde{e}_{j'}(z,\delta) - \widetilde{e}_{k'}(z,\delta))$. Consequently, the domains D and $\overline{D}$ defined above keep the same dissipativity properties when $\delta > 0$ is small enough.

Let us finally turn to the determination of the associated permutation σ. As noticed earlier, the eigenvalues $\widetilde{e}_j(z,\delta)$ are continuous in δ uniformly in $z \in \widetilde{S}_\alpha$. Hence, since the eigenvalues $e_j(z,0)$ are analytic in S_α, we have

$$\lim_{\delta\to 0} \widetilde{e}_j(\infty,\delta) = e_j(\infty,0) \quad j = 1,2,\ldots,n, \tag{6.53}$$

whereas along the real axis (see Figure 4), we have

$$\lim_{\delta\to 0} e_{\sigma(j)}(\infty,\delta) = e_j(\infty,0), \tag{6.54}$$

with σ defined in the lemma, from which the result follows. □

7. Applications. Let us consider the time-dependent Schrödinger equation in the adiabatic limit. The relevant equation is then (2.1), where $H(t) = H^*(t)$ is the time-dependent self-adjoint Hamiltonian. Thus we can take $J = \mathbf{I}$ in Proposition 2.1 to get

$$H(t) = H^*(t) = H^{\#}(t). \tag{7.1}$$

Since the norm of an eigenvector is positive, it remains to impose the gap hypothesis in H3 to fit in the framework, and we deduce that the S-matrix is unitary since $R = \mathbf{I}$. In this context, the elements of the S-matrix describe the transitions between the different levels between $t = -\infty$ and $t = +\infty$ in the adiabatic limit.

We now specify our concern a little further and consider a three-level system, i.e., $H(t) = H^*(t) \in M_3(\mathbf{C})$. We assume that $H(t)$ satisfies the hypotheses of Theorem 6.1 with an extra parameter δ, which we omit in the notation, and displays two avoided crossings at $t_1 < t_2$, as shown in Figure 7. The corresponding permutation σ is given by

$$\sigma(1) = 3, \qquad \sigma(2) = 1, \qquad \sigma(3) = 2. \tag{7.2}$$

By Theorem 6.1, we can compute asymptotically the elements s_{31}, s_{12}, s_{23}, and s_{jj}, $j = 1,2,3$. Using the unitarity of the S-matrix, we can get some more information. Introducing

$$\Gamma_j = \left| \text{Im} \int_{\zeta_j} e_j(z)dz \right|, \quad j = 1,2, \tag{7.3}$$

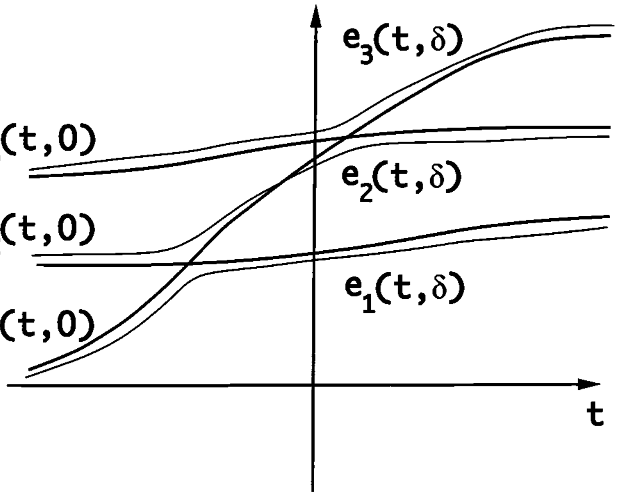


FIG. 7. *The pattern of avoided crossings in the adiabatic context.*

where ζ_j is in the upper half-plane, with the notation of section 6, it follows that

$$s_{31} = \mathcal{O}(e^{-(\Gamma_1+\Gamma_2)/\varepsilon}), \qquad s_{12} = \mathcal{O}(e^{-\Gamma_1/\varepsilon}), \qquad s_{23} = \mathcal{O}(e^{-\Gamma_2/\varepsilon}), \tag{7.4}$$

and

$$s_{jj} = 1 + \mathcal{O}(\varepsilon), \quad j = 1, 2, 3. \tag{7.5}$$

Expressing the fact that the first and second columns as well as the second and third rows are orthogonal, we deduce

$$s_{21} = -\overline{s_{12}}\frac{s_{11}}{\overline{s_{22}}}(1 + \mathcal{O}(e^{-2\Gamma_2/\varepsilon})), \tag{7.6}$$

$$s_{32} = -\overline{s_{23}}\frac{s_{33}}{\overline{s_{22}}}(1 + \mathcal{O}(e^{-2\Gamma_1/\varepsilon})). \tag{7.7}$$

Finally, the estimate in Theorem 6.1 yields

$$s_{13} = \mathcal{O}(\varepsilon e^{-|h|(e_2(\infty,\delta)-e_1(\infty,\delta))/\varepsilon} e^{-\Gamma_2/\varepsilon}) = \mathcal{O}(e^{-(\Gamma_2+\Gamma_2+K)/\varepsilon}), \tag{7.8}$$

where $K > 0$, since we have that $\Gamma_j \to 0$ as $\delta \to 0$. Hence we get

(7.9)
$$S = \begin{pmatrix} s_{11} & s_{12} & \mathcal{O}\left(e^{-(\Gamma_2+\Gamma_2+K)/\varepsilon}\right) \\ -\overline{s_{12}}\frac{s_{11}}{\overline{s_{22}}}\left(1 + \mathcal{O}\left(e^{-2\Gamma_2/\varepsilon}\right)\right) & s_{22} & s_{23} \\ s_{31} & -\overline{s_{23}}\frac{s_{33}}{\overline{s_{22}}}\left(1 + \mathcal{O}\left(e^{-2\Gamma_1/\varepsilon}\right)\right) & s_{33} \end{pmatrix},$$

where all s_{jk}'s above can be computed asymptotically up to exponentially small relative error using (6.10).

The smallest asymptotically computable element s_{31} describes the transition from $e_1(-\infty, \delta)$ to $e_3(+\infty, \delta)$. The result that we obtain for this element is in agreement with the rule of thumb that claims that the transitions take place locally at the avoided crossings and can be considered as independent. Accordingly, we can only estimate the smallest element of all, s_{13}, which describes the transition from $e_3(-\infty,\delta)$ to $e_1(+\infty,\delta)$, for which the avoided crossings are not encountered in the "right order," as discussed in [HP]. It is possible, however, to get an asymptotic expression for this element in some cases. When the unperturbed levels $e_2(z,0)$ and $e_3(z,0)$ possess a degeneracy point in S_α and when there exists a dissipative domain for the index 3 of the unperturbed eigenvalues going above this point, one can convince oneself that s_{13} can be computed asymptotically for δ small enough using the techniques presented above.

Our second application is the study of the semiclassical scattering properties of the multichannel stationary Schrödinger equation with energy above the potential barriers. The relevant equation is then

$$-\varepsilon^2\frac{d^2}{dt^2}\Phi(t)+V(t)\Phi(t)=E\Phi(t), \tag{7.10}$$

where $t\in\mathbf{R}$ is a space variable, $\Phi(t)\in\mathbf{C}^m$ is the wave function, $\varepsilon\to 0$ denotes Planck's constant, $V(t)=V^*(t)\in M_m(\mathbf{C})$ is the matrix of potentials, and the spectral parameter E is kept fixed and large enough so that

$$U(t)\equiv E-V(t)>0. \tag{7.11}$$

Introducing

$$\psi(t)=\begin{pmatrix}\Phi(t)\\ i\varepsilon\Phi(t)\end{pmatrix}\in\mathbf{C}^{2m}, \tag{7.12}$$

we cast equation (7.10) into the equivalent form (2.1) for $\psi(t)$ with the generator

$$H(t)=\begin{pmatrix}\mathbf{O} & \mathbf{I}\\ U(t) & \mathbf{O}\end{pmatrix}\in M_{2m}(\mathbf{C}). \tag{7.13}$$

It is readily verified that

$$H(t)=J^{-1}H^*(t)J, \tag{7.14}$$

with

$$J=\begin{pmatrix}\mathbf{O} & \mathbf{I}\\ \mathbf{I} & \mathbf{O}\end{pmatrix}. \tag{7.15}$$

Concerning the spectrum of $H(t)$, we should remark that if the real and positive eigenvalues of $U(t)$, $k_j^2(t)$, $j=1,\dots,m$ associated with the eigenvectors $u_j(t)\in\mathbf{C}^m$ are assumed to be distinct, i.e.,

$$0<k_1^2(t)<k_2^2(t)<\cdots<k_m^2(t), \tag{7.16}$$

then the spectrum of the generator $H(t)$ given by (7.13) consists of $2m$ real distinct eigenvalues

$$-k_m(t)<-k_{m-1}(t)<\cdots<-k_1(t)<k_1(t)<k_2(t)<\cdots<k_m(t) \tag{7.17}$$

associated with the $2m$ eigenvectors

$$\begin{aligned}\chi_j^{\pm}(t)&=\begin{pmatrix}u_j(t)\\ \pm k_j(t)u_j(t)\end{pmatrix}\in\mathbf{C}^{2m},\\ H(t)\chi_j^{\pm}(t)&=\pm k_j(t)\chi_j^{\pm}(t).\end{aligned} \tag{7.18}$$

We check that

$$(\chi_j^{\pm}(0), \chi_j^{\pm}(0))_J = \pm 2k_j(0)\|u_j(0)\| \neq 0, \quad j = 1, \ldots, m, \tag{7.19}$$

where $\|u_j(0)\|$ is computed in $\mathbf{C}^m$, so that Proposition 2.1 applies. Before dealing with its consequences, we further make explicit the structure of S. Adopting the notation suggested by (7.17) and (7.18), we write

$$H(t) = \sum_{j=1}^{m} k_j(t) P_j^{+}(t) - \sum_{j=1}^{m} k_j(t) P_j^{-}(t), \tag{7.20}$$

$$\psi(t) = \sum_{j=1}^{m} c_j^{+}(t)\varphi_j^{+}(t) \mathrm{e}^{-i\int_0^t k_j(t')dt'/\varepsilon} + \sum_{j=1}^{m} c_j^{-}(t)\varphi_j^{-}(t) \mathrm{e}^{i\int_0^t k_j(t')dt'/\varepsilon} \tag{7.21}$$

and introduce

$$\mathbf{c}^{\pm}(t) = \begin{pmatrix} c_1^{\pm}(t) \\ c_2^{\pm}(t) \\ \vdots \\ c_m^{\pm}(t) \end{pmatrix} \in \mathbf{C}^m. \tag{7.22}$$

Hence we have the block structure

$$S \begin{pmatrix} \mathbf{c}^{+}(-\infty) \\ \mathbf{c}^{-}(-\infty) \end{pmatrix} \equiv \begin{pmatrix} S_{++} & S_{+-} \\ S_{-+} & S_{--} \end{pmatrix} \begin{pmatrix} \mathbf{c}^{+}(-\infty) \\ \mathbf{c}^{-}(-\infty) \end{pmatrix} = \begin{pmatrix} \mathbf{c}^{+}(+\infty) \\ \mathbf{c}^{-}(+\infty) \end{pmatrix}, \tag{7.23}$$

where $S_{\sigma\tau} \in M_m(\mathbf{C})$, $\sigma, \tau \in \{+, -\}$.

Let us turn to the symmetry properties of S. We get from (7.19) and Proposition 2.1 that

(7.24)
$$\begin{pmatrix} S_{++} & S_{+-} \\ S_{-+} & S_{--} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} S_{++} & S_{+-} \\ S_{-+} & S_{--} \end{pmatrix}^{*} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & -\mathbf{I} \end{pmatrix} = \begin{pmatrix} S_{++}^{*} & -S_{-+}^{*} \\ -S_{+-}^{*} & S_{--}^{*} \end{pmatrix}.$$

In terms of the blocks $S_{\sigma\tau}$, this is equivalent to

$$S_{++}S_{++}^{*} - S_{+-}S_{+-}^{*} = \mathbf{I}, \tag{7.25}$$

$$S_{++}S_{-+}^{*} - S_{+-}S_{--}^{*} = \mathbf{O}, \tag{7.26}$$

$$S_{--}S_{--}^{*} - S_{-+}S_{-+}^{*} = \mathbf{I}. \tag{7.27}$$

The block S_{++} describes the transmission coefficients associated with a wave traveling from the right and S_{-+} describes the associated reflexion coefficients. Similarly, S_{--} and S_{+-} are related to the transmission and reflexion coefficients associated with a wave incoming from the left. It should be noted that in the case of equation (7.10), another convention is often used to define an S-matrix (see, for instance, [F1]). This gives rise to a different S-matrix with a similar interpretation. However, it is not difficult to establish a one-to-one correspondence between the two definitions. If the matrix of potentials $V(t)$ is real symmetric, we have further symmetry in the S-matrix.

Lemma 7.1. *Let S given by* (7.23) *be the S-matrix associated with* (7.10) *under condition* (7.11). *If we further assume that $V(t) = \overline{V(t)}$, then taking $\varphi_j^{\pm}(0) \in \mathbf{R}^{2m}$, $j = 1, \ldots, m$, in* (7.21), *we get*

$$S_{++} = \overline{S_{--}}, \qquad S_{+-} = \overline{S_{-+}}. \tag{7.28}$$

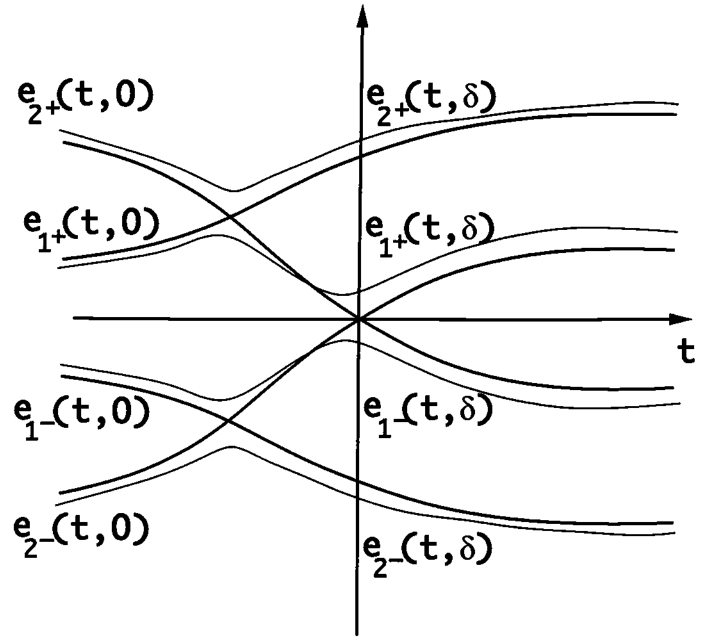


FIG. 8. *The pattern of avoided crossings in the semiclassical context.*

The corresponding results for the S-matrix defined in [F1] are derived in [MN]. The proof of this lemma can be found in Appendix B. We now consider (7.10) in the case where $U(t)=U^*(t)=\overline{U(t)}\in M_2(\mathbf{R})$, which describes a two-channel Schrödinger equation. We assume that the four-level generator $H(t)$ displays three avoided crossings at $t_1<t_2$, two of which take place at the same point t_1 because of the symmetry of the eigenvalues, as in Figure 8. By Lemma 7.1, it is enough to consider the blocks S_{++} and S_{+-}. The transitions corresponding to elements of these blocks which we can compute asymptotically are from level 1^+ to level 2^+ and from level 2^- to level 1^+. They correspond to elements s_{21}^{++} and s_{12}^{+-}, respectively. With the notation

$$\Gamma_j=\left|\operatorname{Im}\int_{\zeta_j}k_1(z)dz\right|,\quad j=1,2, \tag{7.29}$$

where ζ_j is in the upper half-plane, we have the estimates

(7.30)
$$s_{21}^{++}=\mathcal{O}(\mathrm{e}^{-\Gamma_1/\varepsilon}),\qquad s_{12}^{+-}=\mathcal{O}(\mathrm{e}^{-(\Gamma_1+\Gamma_2)/\varepsilon}),\qquad s_{jj}^{++}=1+\mathcal{O}(\varepsilon),\quad j=1,2.$$

It follows from (7.26) and Lemma 7.1 that the matrix $S_{++}S_{+-}^T$ is symmetric. Hence

$$s_{11}^{++}s_{21}^{+-}+s_{12}^{++}s_{22}^{+-}=s_{21}^{++}s_{11}^{+-}+s_{22}^{++}s_{12}^{+-}, \tag{7.31}$$

whereas we get from (7.25) that

$$s_{11}^{++}\overline{s_{21}^{++}}+s_{12}^{++}\overline{s_{22}^{++}}=s_{11}^{+-}\overline{s_{21}^{+-}}+s_{12}^{+-}\overline{s_{22}^{+-}}. \tag{7.32}$$

The only useful estimate we get with Theorem 6.1 is

$$s_{22}^{+-}=\mathcal{O}(\mathrm{e}^{-(\Gamma_1+\Gamma_2+K)/\varepsilon}),\quad K>0, \tag{7.33}$$

which together with (7.30) in (7.31) yields

$$s_{21}^{+-} = s_{21}^{++} s_{11}^{+-} / s_{11}^{++} + \mathcal{O}(\mathrm{e}^{-(\Gamma_1+\Gamma_2)/\varepsilon}). \tag{7.34}$$

Thus from (7.32) and (5.34) for s_{11}^{+-},

$$s_{12}^{++} = -\overline{s_{21}^{++}} \frac{s_{11}^{++}}{\overline{s_{22}^{++}}} (1 + \mathcal{O}(\mathrm{e}^{-\kappa/\varepsilon})), \tag{7.35}$$

with

$$0 < \kappa < \min(\Gamma_1, \Gamma_2). \tag{7.36}$$

Summarizing, we have

$$S_{++} = \begin{pmatrix} s_{11}^{++} & -\overline{s_{21}^{++}} \frac{s_{11}^{++}}{\overline{s_{22}^{++}}} \left(1 + \mathcal{O}\left(\mathrm{e}^{-\kappa/\varepsilon}\right)\right) \\ s_{21}^{++} & s_{22}^{++} \end{pmatrix} \tag{7.37}$$

and

$$S_{+-} = \begin{pmatrix} \mathcal{O}\left(\mathrm{e}^{-\kappa/\varepsilon}\right) & s_{12}^{+-} \\ \mathcal{O}\left(\mathrm{e}^{-\kappa/\varepsilon}\right) & \mathcal{O}\left(\mathrm{e}^{-(\Gamma_1+\Gamma_2+K)/\varepsilon}\right) \end{pmatrix}, \tag{7.38}$$

where all elements $s_{jk}^{\sigma\tau}$ can be asymptotically computed up to exponentially small relative corrections using (6.10). We obtain no information on the first column of S_{+-} except estimate (5.34), where (7.36) necessarily holds. However, if there exists one or several other dissipative domains for certain indices, it is then possible to get asymptotic formulas for the estimated terms.

Appendix A. Proof of Proposition 2.1. A direct consequence of the property

$$H(t) = H^{\#}(t) = J^{-1} H^*(t) J \tag{A.1}$$

is the relation $\sigma(H(t)) = \overline{\sigma(H(t))}$. Thus if $\sigma(H(0)) \subset \mathbf{R}$, then $\sigma(H(t)) \subset \mathbf{R}$ for all $t \in \mathbf{R}$ since the analytic eigenvalues are assumed to be distinct and nondegenerate for all $t \in \mathbf{R}$. Let $e_j(0)$ be the eigenvalue associated with $\varphi_j(0)$. Then due to the property $H(0) = H^{\#}(0)$,

$$(\varphi_j(0), H(0)\varphi_k(0))_J = e_k(0)(\varphi_j(0), \varphi_k(0))_J = \overline{e_j(0)}(\varphi_j(0), \varphi_k(0))_J \tag{A.2}$$

for any $j, k = 1, \ldots, n$. For $j = k$, we get from the assumption $(\varphi_j(0), \varphi_j(0))_J \neq 0$ that $e_j(0) \in \mathbf{R}$, and from the fact that the eigenvalues of $H(0)$ are distinct,

$$(\varphi_j(0), \varphi_k(0))_J = 0, \quad j \neq k. \tag{A.3}$$

The resulting reality of $e_j(t)$ for all $t \in \mathbf{R}$ and $j = 1, \ldots, n$ together with (A.1) yields

$$P_j(t) = J^{-1} P_j^*(t) J. \tag{A.4}$$

Hence using the fact that the P_j^*'s are projectors,

$$\begin{aligned} K(t) = \sum_{j=1}^{n} P_j{}'(t) P_j(t) &= \sum_{j=1}^{n} (J^{-1} P_j^*(t) J)' J^{-1} P_j^*(t) J = J^{-1} \sum_{j=1}^{n} P_j^{*\prime}(t) P_j^*(t) J \\ &= -J^{-1} \sum_{j=1}^{n} P_j^*(t) P_j^{*\prime}(t) J = -J^{-1} K^*(t) J. \end{aligned} \tag{A.5}$$

Let $\Phi, \Psi \in \mathbf{C}^n$ and $W(t)$ be defined by

$$W'(t) = K(t)W(t), \qquad W(0) = \mathbf{I} \tag{A.6}$$

(see (3.5)). Then we have

$$\begin{aligned}(W(t)\Phi, W(t)\Psi)'_J &= \langle W'(t)\Phi | JW(t)\Psi\rangle + \langle W(t)\Phi | JW'(t)\Psi\rangle \\ &= \langle K(t)W(t)\Phi | JW(t)\Psi\rangle + \langle W(t)\Phi | JK(t)W(t)\Psi\rangle \\ &= \langle W(t)\Phi | J(J^{-1}K^*(t)J + K(t))W(t)\Psi\rangle \equiv 0. \end{aligned} \tag{A.7}$$

Thus in the indefinite metric, the scalar products of the eigenvectors of $H(t)$, $\varphi_j(t) = W(t)\varphi_j(0)$ (see (3.7)), are constants:

$$(\varphi_j(t), \varphi_k(t))_J \equiv (\varphi_j(0), \varphi_k(0))_J. \tag{A.8}$$

We can then normalize the $\varphi_j(0)$ in such a way that

$$(\varphi_j(t), \varphi_k(t))_J = (\varphi_j(0), \varphi_k(0))_J = \delta_{jk}\rho_j, \tag{A.9}$$

with $\rho_j \in \{+1, -1\}$. Let $\psi(t)$ and $\chi(t)$ be two solutions of (2.1). By an argument similar to the one above using (A.1), we deduce

$$(\chi(t), \psi(t))_J \equiv (\chi(0), \psi(0))_J. \tag{A.10}$$

Inserting the decompositions

$$\psi(t) = \sum_{j=1}^{n} c_j(t) \mathrm{e}^{-i\int_0^t e_j(t')dt'/\varepsilon} \varphi_j(t), \tag{A.11}$$

$$\chi(t) = \sum_{j=1}^{n} d_j(t) \mathrm{e}^{-i\int_0^t e_j(t')dt'/\varepsilon} \varphi_j(t) \tag{A.12}$$

in this last identity yields

$$\begin{aligned}\sum_{j,k=1}^{n} \overline{d}_k(t) c_j(t) (\varphi_k(t), \varphi_j(t))_J \mathrm{e}^{i\int_0^t (e_k(t') - e_j(t'))/\varepsilon dt'} &= \sum_{j} \overline{d}_j(t)\rho_j c_j(t) \\ \equiv \sum_{j=1}^{n} \overline{d}_j(0)\rho_j c_j(0) &= \sum_{j=1}^{n} \overline{d}_j(\pm\infty)\rho_j c_j(\pm\infty). \end{aligned} \tag{A.13}$$

Since the initial conditions for the coefficients,

$$c_j(-\infty) = \delta_{jk}, \qquad d_j(-\infty) = \delta_{jl}, \tag{A.14}$$

imply

$$c_j(+\infty) = s_{jk}, \qquad d_j(+\infty) = s_{jl}, \tag{A.15}$$

introducing the matrix $R = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_n) \in M_n(\mathbf{C})$, we get from (A.13) that

$$R = S^* R S, \tag{A.16}$$

which is equivalent to the assertion $S^{-1} = RS^*R$. □

Appendix B. Proof of Lemma 7.1. Let $G = G^* = G^{-1}$ be given in block structure by

$$G = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & -\mathbf{I} \end{pmatrix} \in M_{2m}(\mathbf{C}) \tag{B.1}$$

and $H(t)$ be given by (7.13) with $U(t) = \overline{U(t)} = U^*(t)$. Since

$$GH(t)G = -H(t), \qquad \overline{H(t)} = H(t)\,, \tag{B.2}$$

and the eigenvalues of $H(t)$ are real, it is readily verified that

$$GP_j^{\pm}(t)G = P_j^{\mp}(t), \qquad \overline{P_j^{\pm}(t)} = P_j^{\pm}(t), \quad j = 1, \dots, m. \tag{B.3}$$

Hence

$$K(t) = \sum_{\substack{j=1 \\ \tau=\pm}}^{m} P_j^{\tau\prime}(t)P_j^{\tau}(t) = \overline{K(t)} = GK(t)G, \tag{B.4}$$

from which it follows that the solution $W(t)$ of

$$W'(t) = K(t)W(t), \qquad W(0) = \mathbf{I} \tag{B.5}$$

satisfies

$$W(t) = \overline{W(t)} = GW(t)G. \tag{B.6}$$

Since the matrix of potentials $U(0)$ is real symmetric, its eigenvectors $u_j(0)$ may be chosen real so that we can assume that

$$\varphi_j^{\pm}(0) = \begin{pmatrix} u_j(0) \\ \pm k_j(0)u_j(0) \end{pmatrix} \in \mathbf{R}^{2m}. \tag{B.7}$$

Thus it follows from the above that

$$\varphi_j^{\pm}(t) = W(t)\varphi_j^{\pm}(0) \in \mathbf{R}^{2m}, \quad \forall t \in \mathbf{R}, \tag{B.8}$$

and satisfies

$$G\varphi_j^{\pm}(t) = GW(t)GG\varphi_j^{\pm}(0) = W(t)G\varphi_j^{\pm}(0) = \varphi_j^{\mp}(t). \tag{B.9}$$

Finally, the main consequence of (B.2) is that if $\psi(t)$ is a solution of

$$i\varepsilon\psi'(t) = H(t)\psi(t), \tag{B.10}$$

then $\varphi(t) = G\overline{\psi(t)}$ is another solution, as is easily verified. Thus we can write with (7.21), (B.8), and (B.9) that

$$\begin{aligned} \varphi(t) &= \sum_{j=1}^{m} d_j^{+}(t)\varphi_j^{+}(t)\mathrm{e}^{-i\int_0^t k_j(t')dt'/\varepsilon} + \sum_{j=1}^{m} d_j^{-}(t)\varphi_j^{-}(t)\mathrm{e}^{i\int_0^t k_j(t')dt'/\varepsilon} \\ &= \sum_{j=1}^{m} \overline{c_j^{+}(t)}\varphi_j^{-}(t)\mathrm{e}^{i\int_0^t k_j(t')dt'/\varepsilon} + \sum_{j=1}^{m} \overline{c_j^{-}(t)}\varphi_j^{+}(t)\mathrm{e}^{-i\int_0^t k_j(t')dt'/\varepsilon}, \end{aligned} \tag{B.11}$$

i.e.,

$$\begin{aligned} d_j^+(t) &= \overline{c_j^-(t)}, \\ d_j^-(t) &= \overline{c_j^+(t)}, \quad \forall j = 1, \dots, m, \quad \forall t \in \mathbf{R}. \end{aligned} \tag{B.12}$$

Finally, using the definition (7.23) and the above property for $t = \pm\infty$, we get for any $\mathbf{d}^\pm(-\infty) \in \mathbf{C}^m$ that

(B.13)

$$\begin{pmatrix} \mathbf{d}^+(+\infty) \\ \mathbf{d}^-(+\infty) \end{pmatrix} = \begin{pmatrix} S_{++} & S_{+-} \\ S_{-+} & S_{--} \end{pmatrix} \begin{pmatrix} \mathbf{d}^+(-\infty) \\ \mathbf{d}^-(-\infty) \end{pmatrix} = \begin{pmatrix} \overline{S_{--}} & \overline{S_{-+}} \\ \overline{S_{+-}} & \overline{S_{++}} \end{pmatrix} \begin{pmatrix} \mathbf{d}^+(-\infty) \\ \mathbf{d}^-(-\infty) \end{pmatrix},$$

from which the result follows. □

Acknowledgments. It is a great pleasure to thank Charles-Edouard Pfister for many enlightening and fruitful discussions which took place in Marseille and Lausanne. The hospitality of the Institut de Physique Théorique de l'EPFL, where part of this work was performed, is acknowledged.

REFERENCES

[Ba] H. Baklouti, *Asymptotique de largeurs de résonnances pour un modèle d'Effet tunnel microlocal*, thèse, Université de Paris Nord, Paris, 1995.

[Be] M. V. Berry, *Histories of adiabatic quantum transitions*, Proc. Roy. Soc. London Ser. A, 429 (1990), pp. 61–72.

[BE] S. Brundobler and V. Elser, *S-matrix for generalized Landau–Zener problem*, J. Phys. A, 26 (1993), pp. 1211–1227.

[CH1] C. E. Carroll and F. T. Hioe, *Generalization of the Landau–Zener calculation to three-level systems*, J. Phys. A, 19 (1986), pp. 1151–1161.

[CH2] C. E. Carroll and F. T. Hioe, *Transition probabilities for the three-level Landau–Zener model*, J. Phys. A, 19 (1986), pp. 2061–2073.

[D] Yu. N. Demkov, *Adiabatic perturbation of discrete spectrum states*, Soviet Phys. Dokl., 11 (1966), p. 138.

[F1] M. Fedoriuk, *Méthodes Asymptotiques pour les Equations Différentielles Ordinaires Linéaires*, Mir, Moscow, 1987.

[F2] M. V. Fedoryuk, *Analysis* I, in Encyclopaedia of Mathematical Sciences, Vol. 13, R. V. Gamkrelidze, ed., Springer-Verlag, Berlin, New York, Heidelberg, 1989.

[FF] N. Fröman and P. O. Fröman, *JWKB Approximation: Contributions to the Theory*, North–Holland, Amsterdam, 1965.

[HP] J.-T. Hwang and P. Pechukas, *The adiabatic theorem in the complex plane and the semi-classical calculation of non-adiabatic transition amplitudes*, J. Chem. Phys., 67 (1977), pp. 4640–4653.

[J] A. Joye, *Proof of the Landau–Zener formula*, Asymptotic Anal., 9 (1994), pp. 209–258.

[JKP] A. Joye, H. Kunz, and C.-E. Pfister, *Exponential decay and geometric aspect of transition probabilities in the adiabatic limit*, Ann. Phys., 208 (1991), pp. 299–332.

[JP1] A. Joye and C.-E. Pfister, *Exponentially small adiabatic invariant for the Shhrödinger equation*, Comm. Math. Phys., 140 (1991), pp. 15–41.

[JP2] A. Joye and C.-E. Pfister, *Superadiabatic evolution and adiabatic transition probability between two non-degenerate levels isolated in the spectrum*, J. Math. Phys., 34 (1993), pp. 454–479.

[JP3] A. Joye and C.-E. Pfister, *Quantum adiabatic evolution*, in On Three Levels: Micro-, Meso-, and Macro-Approaches in Physics (Leuven Conference Proceedings), M. Fannes, C. Meas, and A. Verbeure, eds., Plenum Press, New York, 1994, pp. 139–148.

[JP4] A. Joye and C.-E. Pfister, *Semi-classical asymptotics beyond all orders for simple scattering systems*, SIAM J. Math. Anal., 26 (1995), pp. 944–977.

[K] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, Heidelberg, 1980.

[Kr] S. G. Krein, *Linear Differential Equations in Banach Spaces*, Transl. Math. Monographs 29, AMS, Providence, RI, 1971.

[M] A. Martinez, *Precise exponential estimates in adiabatic theory*, J. Math. Phys., 35 (1994), pp. 3889–3915.

[MN] P. A. Martin and G. Nenciu, *Semi-classical inelastic S-matrix for one-dimensional N-states systems*, Rev. Math. Phys., 7 (1995), pp. 193–242.

[N] G. Nenciu, *Linear adiabatic theory and applications: Exponential estimates*, Comm. Math. Phys., 152 (1993), pp. 479–496.

[O] F. W. J. Olver, *General connection formulae for Liouville–Green approximations in the complex plane*, Philos. Trans. Roy. Soc. London Ser. A, 289 (1978), pp. 501–548.

[R] T. Ramond, *Semiclassical study of quantum scattering on the line*, preprint, Université de Paris Nord, Paris, 1994.

[Sj] J. Sjöstrand, *Projecteurs adiabatiques du point de vue pseudodifférentiel*, C. R. Acad. Sci. Paris Sér. I Math., 22 (1993), pp. 217–220.

[So] E. A. Solov'ev, *Nonadiabatic transitions in atomic collisions*, Soviet Phys. Uspekhi, 32 (1989), pp. 228–250.

[W] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley–Interscience, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 704–714, May 1997

A CLASSICAL THEOREM ON THE SINGULARITIES OF LEGENDRE SERIES IN C^3 AND ASSOCIATED SYSTEM OF HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS*

PETER A. McCOY[†]

Abstract. A classical theorem of Nehari relates the singularities of Legendre series expansions in C_z with those of associated Taylor's series in C_t. The generalization of Nehari's theorem is known for Legendre series in $C_{z_1 \times z_2}$. In this paper, function theoretic methods develop the analogous relationships between the singularities of series expanded as triple products of Legendre polynomials in $C_{z_1 \times z_2 \times z_3}$ and those of associated analytic functions in C_t. The singularities of these generalized Legendre series are determined by certain elliptic curves. Moreover, these series satisfy a system of hyperbolic partial differential equations (PDEs) in C^3 that are connected to Bochner's study of Poisson processes in R^2.

Key words. Legendre series, singularities, function-theoretic methods

AMS subject classifications. Primary, 35A22, 35A20, 33C45; Secondary, 35C10, 35L99, 30B99

PII. S0036141094273490

1. Introduction.

There is a classical theorem of Szego [15] that relates the singularities of a real zonal harmonic series with the singularities of an analytic function of a single complex variable. Szego's theorem was extended to Legendre series in the complex plane by Nehari [14]. Both theorems relied on the argument used by Hadamard [10] in his multiplication of singularities theorem. Subsequently, Gilbert [7, 8] generalized the Hadamard argument by developing the "envelope method" and used this procedure to study the singularities of harmonic functions in R^3 as well as the singularities of solutions of many classes of elliptic partial differential equations (PDEs). One of the distinct features of the theory is its utility in transforming information about the properties of singularities of analytic functions into the corresponding theorems on the structure of singularities of the solutions. Notable theorems concerning the location and type of the singularities of analytic functions are found in the work of Taylor, Mandelbrojt, and Fabry (see [6, 10]). Examples of their function theoretic extensions are found in the references (see Bergman [4], Begehr and Gilbert [2, 3] and Kracht and Kreyszig [11]).

The Nehari theorem was framed in the setting of hyperbolic PDEs by McCoy [12, 13], who considered analytic functions $F(z_1, z_2)$ that are expanded in series of products of ultraspherical polynomials in the complex space $C^2 := C_{z_1} \times C_{z_2}$. The singularities of these series were shown to lie on a certain quadratic curve embedded in C^2. This curve is defined in relation to the singularities of $F(z_1, z_2)$ and a unique analytic function $f(t)$ of one complex variable known as the "associate." The hyperbolic equation under discussion was used by Bochner [5] to characterize symmetric Poisson

*Received by the editors August 29, 1994; accepted for publication (in revised form) January 26, 1996. This research was supported in part by a grant from the Radar Division, U. S. Naval Research Laboratory, Washington, DC. The U. S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U. S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sima/28-3/27349.html

[†]Mathematics Department, United States Naval Academy, Annapolis, MD 21402-5002 (pam@sma.usna.navy.mil).

processes on the square. We find a generalization of Nehari's theorem to C^3 that can be interpreted as an extension of Bochner's Poisson process equation.

The aim of this study is to extend Nehari's theorem to analytic functions

$$f(z_1, z_2, z_3) = \sum_{n=0}^{\infty} \omega_n a_n P_n(z_1) P_n(z_2) P_n(z_3) \tag{1.1a}$$

of three complex variables $(z_1, z_2, z_3) \in C^3 := C_{z_1} \times C_{z_2} \times C_{z_3}$, where $\omega_n = (n + 1/2)$ and the $P_n(z)$ are the Legendre polynomials of degree "n." Our focus is on the relationship between the singularities of $F(z_1, z_2, z_3)$ and those of the associated analytic function

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \tag{1.1b}$$

of the complex variable $t \in C_t$.

We establish that the singularities of F are located on certain elliptic curves in C^3. These curves are determined by means of function theoretic relationships between the singularities of the function pair $\{F, f\}$. The method is to construct reciprocal integral transforms linking the generalized Legendre series with its associate. Application of the Hadamard, end pinch, and envelope methods provides the appropriate relationships between the singularities of the associated functions.

2. Preliminaries.

We begin by establishing that $F(x_1, x_2, x_3)$ and the associate $f(t)$ are locally analytic. The transforms defining the association will be constructed in due course. Let $f(t)$ be analytic at the origin. Elementary function theory confirms that the radius of convergence of the expansion in (1b), and thus the distance to the first singularity, is $r_o = \lim\sup_{n\to\infty} |a_n|^{-1/n}$. We exclude entire functions and require that $r_0 < \infty$. Similarly, consider the expansion in (1a) for the Legendre series $F(x_1, x_2, x_3)$. The estimate $|P_n(x)| \sim O(1/n^{1/2})$ (see [9, p. 68]) verifies that this series converges in a sphere of radius r_0.

The point is to consider the Legendre series $F(x_1, x_2, x_3)$ in the complex domain. The series is analytically continued to an open set about the origin in $C_{z_1 \times z_2 \times z_3}$. Local analyticity follows by observing that from Darboux's extension of the Laplace–Heine formula [16] one finds the asymptotic estimate $|P_n(z_1)P_n(z_2)P_n(z_3)| \sim \delta^n(z_1)\delta^n(z_2)\delta^n(z_3)$ on the domain $E_{\delta(z_1)} \times E_{\delta(z_2)} \times E_{\delta(z_3)}$. This domain is formed by the tensor product of the ellipses $E_{\delta(z_j)} = \{z_j \in C_{z_j} : |z_j - 1| + |z_j + 1| < \delta(z_j)\}$ for $1 \le j \le 3$ under the assumption that the $\max_{1 \le j \le 3} \delta(z_j)$ is sufficiently small.

The first step is to show that the functions $F(z_1, z_2, z_3)$ and $f(t)$ are associated by integral operators. The kernel for the ascending operator is defined as

$$K(t; z_1, z_2, z_3) = \sum_{n=0}^{\infty} \omega_n t^n P_n(z_1) P_n(z_2) P_n(z_3). \tag{2.1}$$

Following previous arguments, it is easy to check that the kernel is analytic in a neighborhood of the origin in C^4. It follows from Cauchy's theorem that

$$F(z_1, z_2, z_3) = (1/2\pi i) \int_{L_t} f(t) K(t^{-1}; z_1, z_2, z_3) dt/t. \tag{2.2}$$

The contour L_t is a simple closed curve that is homologous to the circle $\gamma : |t| = (\sigma + \epsilon)^{-1}$ modulo the singularities of the integrand, and $\epsilon > 0$ is sufficiently small.

The function element $F(z_1, z_2, z_3)$ defined in (3) by contour deformation is analytic on its domain of association (see Gilbert [8, p. 24]).

The inverse operator is based on the kernel

$$K_*(t; z_1, z_2, z_3) = \sum_{n=0}^{\infty} \omega_n^2 t^n P_n(z_1) P_n(z_2) P_n(z_3), \tag{2.3}$$

which is also analytic in a neighborhood of the origin in C^4. It follows from the orthogonality of the Legendre polynomials $\int_{-1}^{+1} P_n(s) P_m(s) ds = \omega_n^{-1} \delta_{nm}$ that the inversion of (3) is carried out by the transform

$$f(t) = \int_{L_{z_1}} \int_{L_{z_2}} \int_{L_{z_3}} K_*(t; z_1, z_2, z_3) F(z_1, z_2, z_3) dz_1 dz_2 dz_3, \tag{2.4}$$

which is formulated here as a function element on its domain of association. The contours L_{z_1}, L_{z_2}, and L_{z_3} are simple open curves that are homologous to the segment $[-1, +1]$ modulo the singularities of the integrand. The contours have their end points fixed at $\{-1, +1\}$. The function element $f(t)$ associated with $F(z_1, z_2, z_3)$ by (5) is analytic on its domain of association. The kernels in (2) and (4) are related in a simple way as $K_*(t; z_1, z_2, z_3) = [t\partial_t + 1/2] K(t; z_1, z_2, z_3)$ or, more compactly, as

$$K_*(t; z_1, z_2, z_3) = t^{1/2} \partial_t [t^{1/2} K(t; z_1, z_2, z_3)]. \tag{2.5}$$

3. The integral transforms.

The reciprocal integral transforms in (3) and (5) are now reformulated to suit our analysis. We expand the kernel of the ascending operator as a symmetric function of the variables z_1, z_2, and z_3. This symmetric expansion is distinguished by the notation $K(t; z_1, z_2, z_3) := \Gamma(t; z_1, z_2, z_3)$, where

$$\Gamma(t; z_1, z_2, z_3) = \left\{ (1/3) \sum_{j=1}^{3} K_j(t; z_1, z_2, z_3) \right\}. \tag{3.1}$$

The terms in the expansion are

$$K_j(t; z_1, z_2, z_3) = \int_{L_\zeta} K(t; 1, z_j, \zeta) [K(1; z_1, z_2, z_3)]_{z_j = \zeta} d\zeta \tag{3.2}$$

for $1 \leq j \leq 3$. The contour L_t is a simple open curve that is homologous to the segment $[-1, +1]$ modulo the singularities of the integrand and has its end points fixed at $\{-1, +1\}$. Any one of the kernels K_j taken individually would suffice in the role of K but taken alone complicates the analysis later on.

The symmetric expansions of the kernels are now converted to closed-form expressions. Let us examine the first function appearing in the product in the integrand of (8). This function is put into closed form by appealing to Watson's integral [17] as follows:

$$\begin{aligned} K(t; 1, z_j, \zeta) &= (1/4\pi i) \int_{L_\tau} J(t; \Omega(\tau; z_j, \zeta)) d\tau/\tau, \\ J(t; \Omega) &= (1 - t^2)/(1 - 2t\Omega + t^2)^{3/2}, \\ \Omega(\tau; z_j, \zeta) &= z_j \zeta + (1 - z_j^2)^{1/2} (1 - \zeta^2)^{1/2} (\tau + \tau^{-1})/2 \end{aligned} \tag{3.3}$$

for $1 \leq j \leq 3$. The contour L_τ is homologous to the unit circle modulo the singularities of the integrand.

Turning to the second factor in the integrand in (8), we find that the closed-form expressions of these functions are determined by Grosjean [9, p. 68] as $K(1; x_1, x_2, x_3) = \pi^{-1}(1 - x_1^2 - x_2^2 - x_3^2 + 2x_1x_2x_3)^{-1/2}$ for points in the primary domain. The primary domain is defined by the points inside and on the cubic surface $\mathbf{S} : \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 - 2x_1x_2x_3 = 1\}$ that are interior to the cube $[-1, +1]^3$. For points in the cube that are located outside the surface $\mathbf{S}$, the kernel $K(1; x_1, x_2, x_3) = 0$. When the kernel is analytically continued to C^3, the resulting function element

$$K(1; z_1, z_2, z_3) = \pi^{-1}(1 - z_1^2 - z_2^2 - z_3^2 + 2z_1z_2z_3)^{-1/2} \tag{3.4}$$

is understood to originate from the primary domain and the transforms in (8) are referenced to this element. We remark about the behavior of the series $K(1; x_1, x_2, x_3)$ along the sides of the cube. The series converges to a well-defined function off the diagonals $x_2 = \pm x_1$ and exhibits delta function behavior on the diagonals in the plane $x_3 = 1$. These points are not contained in the domain of association.

The kernels in (8) are written in closed form by applying (9) as

$$\begin{aligned} &K_j(t; z_1, z_2, z_3) \\ &= (1/4\pi i) \int_{L_\tau} \int_{L_\zeta} \{(1 - t^2)/(1 - 2t\Omega(\tau; z_j, \zeta) + t^2)^{3/2}\} \\ &\quad \times [K(1; z_1, z_2, z_3)]_{z_j = \zeta} d\zeta\, d\tau/\tau \end{aligned} \tag{3.5}$$

for $1 \leq j \leq 3$. This step converts the kernel in (7) into the closed symmetric expression

$$\begin{aligned} &\Gamma(t; z_1, z_2, z_3) \\ &= (1/12\pi i) \sum_{j=1}^{3} \left\{ \int_{L_\tau} \int_{L_\zeta} J(t; \Omega(\tau; z_j, \zeta))[K(1; z_1, z_2, z_3)]_{z_j = \zeta} \right\} d\zeta\, d\tau/\tau. \end{aligned} \tag{3.6}$$

The kernel of the inverse operator is reformulated by a similar procedure. Define

$$\Gamma_*(t; z_1, z_2, z_3) = \left\{ (1/3) \sum_{j=1}^{3} K_{j,*}(t; z_1, z_2, z_3) \right\}, \tag{3.7}$$

where the $K_{j,*}(t; z_1, z_2, z_3) = \int_{L_\zeta} K_*(t; 1, z_j, \zeta)[K(1; z_1, z_2, z_3)]_{z_j = \zeta} d\zeta$, and rely on (6) to write

$$K_*(t; 1, z_j, \zeta) = (1/4\pi i) \int_{L_\tau} t^{1/2} \partial_t [t^{1/2} J(t; \Omega(\tau; z_j, \zeta))]\, d\tau/\tau \tag{3.8}$$

for $1 \leq j \leq 3$. The kernel of the inverse operator is now put into closed symmetric form as

$$\Gamma_*(t; z_1, z_2, z_3) = t^{1/2} \partial_t [t^{1/2} \Gamma(t; z_1, z_2, z_3)]. \tag{3.9}$$

The principal branches of the function elements are taken. We note that $\Gamma(t^{-1}; z_1, z_2, z_3) = -t\Gamma(t; z_1, z_2, z_3)$. This observation brings us to the integral transform pair whose kernels are both in closed form and invariant under cyclic permutations of the variables z_1, z_2, and z_3.

Theorem 3.1. *The T-transforms linking the function elements F and f on their domains of association are*

$$F(z_1, z_2, z_3) = T[f] := (-1/2\pi i) \int_{L_t} f(t)\Gamma(t; z_1, z_2, z_3)dt, \tag{16a}$$

$$f(t) = T^{-1}[F] := \int_{L_{z_1}} \int_{L_{z_2}} \int_{L_{z_3}} \Gamma_*(t; z_1, z_2, z_3)F(z_1, z_2, z_3)dz_1dz_2dz_3. \tag{16b}$$

4. The singular manifolds. The function elements in (16) are analytically continued from their initial domains of definition to larger domains of association by contour deformation. During the continuation, the singularities in the integrands move, and the contours are deformed to avoid singularities that approach them along intersecting trajectories. There may be circumstances under which no further deformation is possible without contact occurring between the contour and the singularity at a particular point. Such a point is a singularity of the integrand but is not necessarily a singularity of the transformed function element. The object is to extract the actual singularities of a function element from the set of possible singularities. An authentic singularity is identified if it corresponds to a singularity of the associated element under the inverse transform. There are three types of circumstances in which this event could occur. The corresponding singularities are referred to as the Hadamard, end pinch, and envelope singularities.

Let us begin the discussion by considering a function element $F(z_1, z_2, z_3)$ whose associate $f(t)$ has an isolated singularity at the point $t = \alpha$. The set of possible singularities of $F(z_1, z_2, z_3)$ is expressed as $PS[F] = [SH(f; \alpha)] \cup [EP(f; \alpha)] \cup [SE(f; \alpha)]$, where the constituent sets are the respective (possible) Hadamard, end pinch, and envelope singularities. The set of possible singularities $PS[F] := PM(f; \alpha)$ is determined from the integrand of the transform in (16a) as having the parametric form

$$PM(f; \alpha) = \cup_{1\leq j\leq 3}\{PM_j(f; \alpha) \cap B_j(Z)\}, \tag{4.17}$$

where the sets

$$PM_j(f; \alpha) = \{Z : \sigma - \Omega(\tau; z_j, \zeta) = 0; (\tau, \zeta) \in L_\tau \times L_\zeta, \varpi(\alpha, \sigma) = \varpi(\tau, \eta) = 0\}$$

and

$$B_j(Z) = \{Z : \varphi(Z) = 0; z_j = \zeta, \zeta \in L_\zeta\}, \quad 1 \leq j \leq 3.$$

For convenience, we designate $Z := (z_1, z_2, z_3)$ or $(Z) := (z_1, z_2, z_3)$, where the interpretation is clear in context. The auxiliary functions are defined as $\varphi(x, y, z) := 1 - x^2 - y^2 - z^2 + 2xyz = (1 - y^2)(1 - z^2) - (x - yz)^2$ and $\varpi(x, y) := x^2 - 2xy + 1$. The function φ is invariant under cyclic permutations of its argument; i.e., $\varphi(x, y, z) = \varphi(y, z, x) = \varphi(z, x, y)$. The analysis will be facilitated by rewriting the sets

$$PM_j(f; \alpha) = \{Z : \psi(\sigma, z_j, \zeta, \eta) = 0, (\tau, \zeta) \in L_\tau \times L_\zeta; \varpi(\alpha, \sigma) = \varpi(\tau, \eta) = 0\},$$
$$1 \leq j \leq 3,$$

where the function $\psi(\sigma, z, \zeta, \eta) := (\sigma - z\zeta)^2 - (1 - z^2)(1 - \zeta^2)\eta^2$.

The set of Hadamard singularities $SH(f; \alpha) = \cup_{1\leq j\leq 3}SH_j(f; \alpha)$ corresponding to the Hadamard singularities $SH_j(f; \alpha)$ contained in the sets $PM_j(f; \alpha) \cap B_j(Z)$ is obtained by eliminating the parameter "ζ" from the principal branches of the manifolds. Working with the set $B_j(Z)$ defined in (17) produces roots $\zeta = \xi_\pm(z_2, z_3) := z_2 z_3 \pm (1 - z_2^2)^{1/2}(1 - z_3^2)^{1/2}$ for the case $j = 1$. The eliminants of the remaining sets $B_2(Z)$ and $B_3(Z)$ are also computed. The end result is that the Hadamard singularities for the ascending operator in (16a) are defined parametrically by the surfaces

$$PH(f; \alpha) = M((z_1, z_2, z_3); \alpha) \cup M((z_2, z_3, z_1); \alpha) \cup M((z_3, z_1, z_2); \alpha), \tag{4.18}$$

where

$$M((x, y, z); \alpha) = \{(x, y, z) \in C^3 : \psi(\sigma, x, \zeta, \eta) = 0, \zeta = \xi_\pm(y, z); \varpi(\alpha, \sigma) = \varpi(\tau, \eta) = 0\}.$$

The next step is to consider the set of possible singularities for the function element $f(t)$ defined by the descending operator in (16b). Designate this set as $PS[f] = [SH(F; Z_o)] \cup [EP(F; Z_o)] \cup [SE(F; Z_0)]$, where $SH(F; Z_o)$, $EP(F; Z_o)$, and $SE(F; Z_o)$ are the respective Hadamard, end pinch, and envelope singularities. Let the associate $F(z_1, z_2, z_3)$ be singular at $Z_o = (z_{1,o}, z_{2,o}, z_{3,o}) \in B(Z)$, where $B(Z) = \{Z : \varphi(Z) = 0\}$. By examining the T-transforms and taking advantage of the role played by the kernel $\Gamma(t; Z)$ in the integrands, the relevant sets are determined to be

$$PM(F; Z_o) = \cup_{1 \le j \le 3} PM_j(F; Z_o), \tag{4.19}$$

where the sets

$$\begin{aligned} PM_j(F; Z_o) = \{t \in C : \sigma - \Omega(\tau; z_j; \zeta) = 0; (\tau, \zeta, Z) \in L_\tau \times L_\zeta \times L_{z_1} \times L_{z_2} \times L_{z_3}; \\ Z_0 = (z_{1,o}, z_{2,o}, z_{3,o}) \in B_j(Z); \varpi(t, \sigma) = \varpi(\tau, \eta) = 0\} \end{aligned}$$

are defined parametrically for $1 \le j \le 3$.

The set of Hadamard singularities of the function element $f(t)$ is $SH(F; Z_o) = \cup_{1 \le j \le 3} SH_j(F; Z_o)$. The singularities contained in this set are embedded in the set $PM(F; Z_o)$. By eliminating the parameters, one determines that the singularities of the descending operator defined in (16b) are

$$SH(F; Z_o) = M_*(t; (z_{1,o}, z_{2,o}, z_{3,o})) \cup M_*(t; (z_{2,o}, z_{3,o}, z_{1,o})) \cup M_*(t; (z_{3,o}, z_{1,o}, z_{2,o})), \tag{4.20}$$

where

$$M_*(t; (x, y, z)) = \{t \in C : \varpi(t, \sigma) = \varpi(\tau, \eta) = 0; \psi(\sigma, x, \zeta, \eta) = 0, \zeta = \xi_\pm(y, z)\}.$$

We observe that the preceding arguments are symmetric. Bearing in mind that the actual singularities determined in the analysis are, in general, subsets of the sets of Hadamard singularities of the associated function pair $\{F, f\}$, we summarize the results in the following statement.

THEOREM 4.1. *Let $\{F, f\}$ be a T-transform associated function element pair. Then the points Z_o and t_o are the respective Hadamard singularities if and only if $Z_o \in SH(f; t_o)$ and $t_o \in SH(F; Z_o)$.*

It is an easy matter at this point to dispose of the end pinch singularities. These singularities are distinguished by the fixed points $\{-1, +1\}$ at the ends of open contours of integration. As end points, they remain fixed during the contour deformation and, thus, must be considered in terms of singularities that approach them. The end pinch singularities of the ascending operator are expressed in parametric form as

$$EP(f;\alpha) = \cup_{1\le j\le 3}\{EP_j(f;\alpha) \cap EB_j(f;\alpha)\}.$$

The sets of singularities

$$\begin{aligned} EP_j(f;\alpha) = \{Z; \sigma - \Omega(\tau; z_j, \pm 1) = 0,\\ \partial_\tau[\sigma - \Omega(\tau; z_j, \pm 1)] = 0, \tau \in L_\tau; \varpi(\alpha,\sigma) = \varpi(\tau,\eta) = 0\}\end{aligned}$$

are of a "mixed" type since L_τ is a closed contour. The reasoning requires a combination of end pinch and envelope arguments which leads to the eliminant

$$EP_j(f;\alpha) = \{Z : \varphi(Z) = 0, z_j = \pm\sigma;\ \varpi(z,\sigma) = 0\}.$$

The sets $EB_j(Z) = \{Z : \varphi(Z) = 0,\ z_j = \zeta = \pm\sigma\}$ for $1 \le j \le 3$. Combining these sets yields the following for $1 \le j \le 3$:

$$EP_j(f;\alpha) \cap EB_j(f;\alpha) = \{Z : \varphi(Z) = 0, z_j = \pm\sigma;\ \varpi(\alpha,\sigma) = 0\}. \tag{4.21a}$$

Similarly, the end pinch singularities of the descending operator are

$$EP(F; Z_o) = \cup_{1\le j\le 3} EP_j(F; Z_o),$$

where

$$\begin{aligned} EP_j(F;Z_o) = \{t \in C : \sigma - \Omega(\tau; z_j, \pm 1) = \partial_\tau[\sigma - \Omega(\tau; z_j, \pm 1)] = 0;\\ \tau \in L_\tau, z_j = \pm 1; Z \in EB_j(F; Z_o), \varpi(t,\sigma) = \varpi(\tau,\eta) = 0\}\end{aligned}$$

for $1 \le j \le 3$. In the sets $EB_j(F;Z_o) = \{Z : (Z) = (Z_o) = (\pm 1, \pm 1, \pm 1);\ \varphi(Z) = 0\}$ the minus signs are selected pairwise. The eliminant in this case is

$$EP_j(F;Z_o) = \{t \in C_t : \varphi(Z) = 0, z_j = \pm\sigma \text{ and } z_k = \pm 1 \text{ for } z_k \ne z_j; \varpi(t,\sigma) = 0\}. \tag{4.21b}$$

We state the following result to summarize the analysis leading to (21).

THEOREM 4.2. *Let $\{F, f\}$ be a T-transform associated function pair. If $f(t)$ is singular at the point t_o, then $EP(f;t_o)$ is the set of end pinch singularities of $F = T[f]$. In addition, if $F(Z)$ is singular at $Z_o : \varphi(Z_o) = 0$, the set of end pinch singularities of the element function $f = T^{-1}[F]$ is $EP(F;Z_o)$.*

We now turn to the envelope singularities of the transform pair. The envelope singularities of the ascending operator are

$$SE(f;\alpha) = \cup_{1\le j\le 3}\{SE_j(f;\alpha) \cap SE_j^*(f;\alpha)\},$$

where the sets are defined in terms of the auxiliary functions

$$\Phi_j(\sigma,\tau,Z,\zeta) = [\sigma - \Omega(\tau; z_j, \zeta)][\varphi(z_1,z_2,z_3)|_{z_j=\zeta}],$$

since

$$SE(f;\alpha) = \{Z : \Phi_j(\sigma,\tau,Z,\zeta) = 0, (\tau,\zeta) \in L_\tau \times L_\zeta;\ \varpi(\alpha,\sigma) = \varpi(\tau,\eta) = 0\}$$

and

$$\begin{aligned} SE_j^*(f;\alpha) = \ &\{Z : \partial_\tau\Phi_j(\sigma,\tau,Z,\zeta) = \partial_\zeta\Phi_j(\sigma,\tau,Z,\zeta) = 0,\\ &(\tau,\zeta) \in L_\tau \times L_\zeta;\ \varpi(\alpha,\sigma) = \varpi(\tau,\eta) = 0\}, \quad 1 \le j \le 3.\end{aligned}$$

We will eliminate the variables from the three systems

$$\Phi_j(\sigma,\tau,Z,\zeta)=0, \tag{4.22a}$$

$$\partial_\tau\Phi_j(\sigma,\tau,Z,\zeta)=0, \tag{4.22b}$$

$$\partial_\zeta\Phi_j(\sigma,\tau,Z,\zeta)=0, \tag{4.22c}$$

$1\le j\le 3$. Working with (22a,b) and eliminating "τ" give the equations $\sigma-\Omega(\tau;z_j;\zeta)|_{\tau=\pm1}=0$, which lead to the parametric form

$$SE_{1,1}(f;\alpha)=\bigcup_{1\le j\le 3}\{Z:\varphi(\sigma,z_j,\zeta)=0,\varphi(Z)|_{z_j=\zeta}=0;\ \varpi(\alpha,\sigma)=0\} \tag{4.23a}$$

when $\tau=1$. Solving the equation $\varphi(z_1,z_2,z_3)=0$ for the variable "z_1" yields $z_1=\xi_\pm(z_2,z_3)$ with $z_1=\zeta$ as expected. We do not place this into (23a) but leave (23a) in its parametric form. The other cases are analogous. The choice $\tau=-1$ results in $\sigma=z_j\zeta$ and $\varphi(z_1,z_2,z_3)|_{z_j=\zeta}=0$ for $1\le j\le 3$. The remaining cases are similar, and one eliminates "ζ" in each pair of equations to find that

$$SE_{1,2}(f;\alpha)=\bigcup_{1\le j\le 3}\{Z:\varphi(Z)|_{z_j\to\sigma/z_j}=0,z_j\ne 0;\ \varpi(\alpha,\sigma)=0\}. \tag{4.23b}$$

The arrow symbol ($\to$) is read as "replace by." We combine these results as

$$SE_1^*(f;\alpha)=SE_{1,1}(f;\alpha)\cup SE_{1,2}(f;\alpha).$$

Turning our attention to (22a,c), we find that

$$\partial_\zeta\Phi_j(\sigma,\tau,Z,\zeta)=[\sigma-\Omega(\tau;z_j;\zeta)]\partial_\zeta\varphi(Z)|_{z_j=\zeta}+\partial_\zeta[\sigma-\Omega(\tau;z_j;\zeta)]\varphi(Z)|_{z_j=\zeta}=0$$

for $1\le j\le 3$. Evaluating the equation $\partial_\zeta\Phi_j=0$ under the presumption that $\varphi(Z)|_{z_j=\zeta}=0$ gives $[\sigma-\Omega(\tau;z_j;\zeta)](-2\zeta+2z_kz_l)=0$, where $\{k,l\}\in\{1,2,3\}\backslash\{j\}$, $k\ne l$. The eliminants produce

$$\begin{aligned}SE_{2,1}(f;\alpha)=\bigcup_{1\le j\le 3}\{Z:&\ \psi(\sigma,z_j,\zeta,\eta)|_{\zeta\to z_kz_l}=0;\\ &\varphi(z_j,z_k,z_l)|_{z_j\to z_kz_l}=0;\ \varpi(\alpha,\sigma)=\varpi(\tau,\eta)=0\},\end{aligned} \tag{4.23c}$$

where the indices $\{j,k,l\}$ are distinct. On the other hand, proceeding from $\partial_\zeta\Phi_j=0$ (under the assumption that $\sigma-\Omega(\tau;z_j;\zeta)=0$) yields $z_j^2(1-\zeta^2)=(1-z_j^2)\zeta^2\eta^2$ and $\zeta^2=\kappa(z_j,\eta)$, where $\kappa(x,\eta):=x^2/(\eta^2+(1-\eta^2)x^2)$. The resulting sets are stated parametrically as

$$SE_{2,2}(f;\alpha)=\bigcup_{1\le j\le 3}\{Z:\sigma-\Omega(\tau;z_j;\zeta)=0,\zeta^2=\kappa(z_j,\eta);\ \varpi(\alpha,\sigma)=\varpi(\tau,\eta)=0\}. \tag{4.23d}$$

The other cases involving $\partial_\zeta\Phi_j=0$ are part of previous discussions. We define

$$SE_2^*(f;\alpha)=SE_{2,1}(f;\alpha)\cup SE_{2,2}(f;\alpha).$$

In summary, the envelope singularities of $F=T[f]$ that do not reduce to Hadamard or end pinch singularities are given by

$$SE(f;\alpha)=SE_1^*(f;\alpha)\cup SE_2^*(f;\alpha).$$

We consider the singularities for the descending operator $f = T^{-1}[F]$, focusing on singularities that were not previously identified as of the Hadamard or end pinch type. We define the sets

$$SE(F; Z_o) = \cup_{1\leq j\leq 3}\{SE_j(F; Z_o) \cap SE_j^*(F; Z_o)\},$$

where

$$SE_j(F; Z_o) = \{t \in C_t : \Phi_j(\sigma, \tau, Z, \zeta) = 0, (\tau, \zeta, Z) \in L_\tau \times L_\zeta \times L_{z_1} \times L_{z_2} \times L_{z_3}; \\ \varpi(t, \sigma) = \varpi(\tau, \eta) = 0\}$$

and

$$SE_j^*(F; Z_o) = \{t \in C_t : \partial_\tau\Phi_j(\sigma, \tau, Z, \zeta) = \partial_\zeta\Phi_j(\sigma, \tau, Z, \zeta) = \partial_{z_k}\Phi_j(\sigma, \tau, Z, \zeta) = 0, \\ 1 \leq k \leq 3; \\ (\tau, \zeta, Z) \in L_\tau \times L_\zeta \times L_{z_1} \times L_{z_2} \times L_{z_3}; \varpi(t, \sigma) = \varpi(\tau, \eta) = 0\}$$

for $1 \leq j \leq 3$. The relevant systems of equations are

$$\Phi_j(\sigma, \tau, Z, \zeta) = 0, \tag{4.24a}$$

$$\partial_\tau\Phi_j(\sigma, \tau, Z, \zeta) = \partial_\zeta\Phi_j(\sigma, \tau, Z, \zeta) = 0, \tag{4.24b}$$

$$\partial_{z_k}\Phi_j(\sigma, \tau, Z, \zeta) = 0 \tag{4.24c}$$

for $1 \leq j, k \leq 3$. The eliminants for (24a,b) are analogous with those for (22), (23). The results are

$$SE_1^*[F; Z_o] = SE_{1,1}[F; Z_o] \cup SE_{1,2}[F; Z_o],$$

where

$$SE_{1,1}[F; Z_o] = \bigcup_{1\leq j\leq 3}\{t \in C_t : \varpi(t, \sigma) = 0; \varphi(Z_o)|_{z_{o,j}=\zeta} = 0, \varphi(\sigma, z_{o,j}, \zeta) = 0, \varphi(Z_o) = 0\}$$

and

$$SE_{1,2}[F; Z_o] = \bigcup_{1\leq j\leq 3}\{t \in C_t : \varpi(t, \sigma) = 0; \varphi(Z_o)|_{z_{o,j}\to\sigma/z_{o,j}} = 0, \varphi(Z_o) = 0\}.$$

In similar fashion, we find the sets

$$SE_2^*[F; Z_o] = SE_{2,1}[F; Z_o] \cup SE_{2,2}[F; Z_o],$$

where

$$SE_{2,1}[F; Z_o] = \bigcup_{1\leq j\leq 3}\{t \in C_t : \varpi(t, \sigma) = 0; \psi(\sigma, z_{o,j}, \zeta, \eta)|_{\zeta=z_{o,k}z_{o,l}} = 0, \\ \varphi(Z_o) = 0, \varpi(\tau, \eta) = 0\}$$

and

$$SE_{2,2}[F; Z_o] = \bigcup_{1\leq j\leq 3}\{t \in C_t : \varpi(t, \sigma) = 0; \sigma - \Omega(\tau, z_{o,j}, \zeta) = 0, \\ \zeta^2 = \kappa(z_{o,j}, \eta), \varphi(Z_o) = 0,\ \varpi(\tau, \eta) = 0\}.$$

By symmetry, the eliminants of (24a,c) are identical with those of (22a,c) in the case $j = k$. Let us consider the remaining eliminants in (24c) for $j \neq k$. One computes here that $\partial_{z_k}\Phi_j(\sigma, \tau, Z, \zeta) = 0$ and finds that

$$SE_3^*[F; Z_o] = \bigcup_{1\le j\le 3} \{t \in C_t : \varpi(t, \sigma) = 0; \psi(\sigma, z_j, \zeta, \eta) = 0,$$
$$\partial_{z_k}\varphi|_{z_j=\zeta} = 0, \partial_{z_l}\varphi|_{z_j=\zeta} = 0, \varphi(Z_o) = 0, \varpi(\tau, \eta) = 0\},$$

where the indices are taken as distinct. This set has been specified in parametric form with "ζ" defined implicitly for brevity. By combining the three cases, the envelope set of the descending operator is thus

$$SE[F; Z_o] = \bigcup_{1\le j\le 3} SE_j^*[F; Z_o].$$

We summarize our findings as follows.

THEOREM 4.3. *The set of envelope singularities of the function element $F(Z)$ represented by the transform $F = T[f]$ is $SE(f; t_o)$, where the associate $f(t)$ is singular at the point t_o. The set of envelope singularities of the function element $f(t)$ represented by the transform $f = T^{-1}[F]$ is $SE(F; Z_o)$, where the associate $F(Z)$ is singular at $Z_o : \varphi(Z_o) = 0$. Moreover, the point $Z_o \in SE(f; t_o)$ is a true envelope singularity of $F(Z)$ if and only if $t_o \in SE(F; Z_o)$ is the corresponding singularity of the function element $f(t)$ determined by $f = T^{-1}[F]$.*

5. Hyperbolic PDEs and Poisson processes. Function theory typically studies singularities of elliptic PDEs in terms of analytic functions. An interesting property of the generalized Legendre series F is that they are solutions of a system of hyperbolic PDEs in C^3 that form an extension of a problem studied by Bochner [5] in connection with Poisson processes in R^2. Bochner views the hyperbolic equation $\{\partial_{x_1}[\sigma_{\alpha,\beta}(x_1, x_2)\partial_{x_1}] - \partial_{x_2}[\sigma_{\alpha,\beta}(x_2, x_1)\partial_{x_2}]\}F(x_1, x_2) = 0, \sigma_{\alpha,\beta}(x_1, x_2) = (1 - x_1)^{\alpha+1}(1 + x_1)^{\beta+1}(1 - x_2)^{\alpha}(1 + x_2)^{\beta}$ on the square $[-1, +1]^2$ for parameters α, $\beta \ge -1/2$, $\alpha + \beta \ge -1/2$. The boundary condition $F(x_1, 1) = g(x_1)$ is imposed along a side $-1 \le x_1 \le +1$.

Consider the system of hyperbolic PDEs

$$\{\partial_{z_1}[\rho(z_1)\partial_{z_1}] - \partial_{z_2}[\rho(z_2)\partial_{z_2}]\}F(z_1, z_2, z_3) = 0,$$
$$\{\partial_{z_2}[\rho(z_2)\partial_{z_2}] - \partial_{z_3}[\rho(z_3)\partial_{z_3}]\}F(z_1, z_2, z_3) = 0,$$
$$\{\partial_{z_3}[\rho(z_3)\partial_{z_3}] - \partial_{z_1}[\rho(z_1)\partial_{z_1}]\}F(z_1, z_2, z_3) = 0$$

for $(z_1, z_2, z_3) \in C^3$ with $\rho(z) = \sigma_{0,0}(z, z)$. In R^3 it is easy to use Legendre's equation to verify that the series $F(x_1, x_2, x_3)$ is a solution of the system in a neighborhood of the origin and that it satisfies the boundary conditions of the type $F(x_1, 1, 1) = g(x_1)$, $-1 \le x_1 \le +1$ on the faces of the cube $[-1, +1]^3$. The series $F(z_1, z_2, z_3)$ admits an analytic continuation to C^3 which may be considered in the context of Bochner's problem as a process that has been extended to several complex variables. The curves and surfaces containing the singularities of the solutions may be interpreted as locating a distribution that creates the process.

Acknowledgement. The author thanks the referee for the constructive commentary.

REFERENCES

[1] W. N. Bailey, *The generating function of Jacobi polynomials*, London Math. Soc. J., 13 (1938), pp. 8–12.

[2] H. Begehr and R. P. Gilbert, *Transformations, Transmutations, and Kernel Functions*, Vol. 1, Pitman Monographs and Surveys in Pure and Applied Mathematics 58, Longman Scientific & Technical, Harlow; John Wiley, New York, 1992.

[3] H. Begehr and R. P. Gilbert, *Transformations, Transmutations, and Kernel Functions*, Vol. 2, Pitman Monographs and Surveys in Pure and Applied Mathematics 58, Longman Scientific and Technical/John Wiley, Harlow, UK/New York, 1992.

[4] S. Bergman, *Integral Operators in the Theory of Linear Partial Differential Equations*, Ergebnisse der Mathematik und ihrer Grenzgebiete 23, Springer-Verlag, New York, 1969.

[5] S. Bochner, *Sturm–Liouville and heat equations whose eigenfunctions are ultraspherical polynomials or associated Bessel functions*, in Proc. Conference on Differential Equations, University of Maryland, College Park, MD, 1956, pp. 23–48.

[6] P. Dienes, *The Taylor's Series*, Dover, New York, 1957.

[7] R. P. Gilbert, *Singularities of three-dimensional harmonic functions*, Pacific J. Math., 10 (1960), pp. 1243–1255.

[8] R. P. Gilbert, *Function Theoretic Methods in Partial Differential Equations*, Mathematics in Science and Engineering, 54, Academic Press, New York, 1969.

[9] C. C. Grosjean, *On the detection and calculation of some remarkable integrals and related expressions involving classical orthogonal polynomials, Part* II, Simon Stevin, 6 (1993), pp. 49–106.

[10] E. Hille, *Analytic Function Theory*, Vols. 1 and 2, Blaisdell, New York, 1962.

[11] M. Kracht and E. O. Kreyszig, *Methods of Complex Analysis in Partial Differential Equations with Applications*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley, New York, 1988.

[12] P. A. McCoy, *Singularities of hyperbolic PDE's in two complex variables*, Comput. Math. Appl., 12A (1986), pp. 551–556.

[13] P. A. McCoy, *Singularities of Jacobi series on C^2 and the Poisson process equation*, J. Math. Anal. Appl., 128 (1987), pp. 92–100.

[14] Z. Nehari, *On the singularities of Legendre expansions*, J. Rational Mech. Anal., 5 (1956), pp. 987–992.

[15] G. Szego, *On the singularities of zonal harmonic expansions*, J. Rational Mech. Anal., 3 (1954), pp. 561–564.

[16] G. Szego, *Orthogonal Polynomials*, 3rd ed., Colloquium Publications, Vol. 23, AMS, Providence, RI, 1967.

[17] G. N. Watson, *Notes on generating functions of polynomials* 3: *Polynomials of Legendre and Gegenbauer*, London Math. Soc. J., 8 (1933), pp. 289–292.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 715–730, May 1997

012

ON TRIGONOMETRIC SERIES EXPANSIONS OF TWELVE JACOBIAN ELLIPTIC FUNCTIONS[*]

D. S. TSELNIK[†]

Abstract. Trigonometric series expansions of the Jacobian elliptic functions *snu*, *cnu*, and *dnu*, which are intermediates between the expansions of these functions that are known in the literature, are derived and discussed. Similar expansions for the functions *cdu*, *sdu*, . . . , *scu* are also derived. From these new expansions, a number of interesting infinite series can be obtained. (Some examples are given.) The usage of these new expansions in applications is discussed.

Key words. Jacobian elliptic functions, trigonometric series expansions

AMS subject classifications. 33E05, 42A16

PII. S0036141095291257

1. Introduction.

In this paper, we consider trigonometric series expansions of (twelve) Jacobian elliptic functions [3], [4], [5], [12], $sn(u,k)$, $cn(u,k)$, $dn(u,k)$, $cd(u,k)$, $sd(u,k),\ldots$, and $sc(u,k)$.

We shall start from the expansions of the three "base" functions *snu*, *cnu*, and *dnu*, giving the pertinent derivations with sufficient details. Then we shall present the expansions of the other nine functions in question—without derivations but with an explanation of how these expansions were obtained.

Therefore, let us consider the functions $sn(u,k)$, $cn(u,k)$, and $dn(u,k)$. Each of these is a function of the complex variable u and the (complex) parameter k (called modulus) [12, section 22.11]. In the u-plane, each of the functions *snu*, *cnu*, and *dnu* has poles at points

$$u = 2nK + (2m+1)iK', \tag{1.1}$$

where $n = 0, \pm 1, \pm 2, \ldots, m = 0, \pm 1, \pm 2, \ldots$, and K and K' are the complete and the associated complete elliptic integrals of the first kind [12, section 22.35].

In the literature, the following trigonometric series expansions of $sn(u,k)$, $cn(u,k)$, and $dn(u,k)$ are known:

$$sn(u,k) = \frac{2\pi}{kK}\sum_{n=0}^{\infty}\frac{q^{n+\frac{1}{2}}}{1-q^{2n+1}}\sin\frac{(2n+1)\pi u}{2K}, \tag{1.2}$$

$$cn(u,k) = \frac{2\pi}{kK}\sum_{n=0}^{\infty}\frac{q^{n+\frac{1}{2}}}{1+q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K}, \tag{1.3}$$

$$dn(u,k) = \frac{\pi}{2K} + \frac{2\pi}{K}\sum_{n=0}^{\infty}\frac{q^{n+1}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K}, \tag{1.4}$$

and

$$sn(u,k) = \frac{\pi}{2kK}\sum_{j=-\infty}^{+\infty}\operatorname{cosec}\frac{\pi}{2K}[u-(2j-1)iK'], \tag{1.5}$$

[*] Received by the editors September 5, 1995; accepted for publication (in revised form) February 14, 1996.
http://www.siam.org/journals/sima/28-3/29125.html

[†] 2416 18th Street South, Apartment 204, Fargo, ND 58103.

$$cn(u,k) = \frac{\pi i}{2kK} \sum_{j=-\infty}^{+\infty} (-1)^j \ \text{cosec}\ \frac{\pi}{2K}[u-(2j-1)iK'], \tag{1.6}$$

$$dn(u,k) = \frac{\pi i}{2K} \lim_{m\to\infty} \sum_{j=-m}^{m} (-1)^j \cot \frac{\pi}{2K}[u-(2j-1)iK'] \tag{1.7}$$

[12, section 22.6], [3, p. 260], [4, sections 8.7 and 8.8].

Here

$$q = \exp(\pi i \tau), \quad \tau = \frac{iK'}{K}, \tag{1.8}$$

and $\text{Im}\tau > 0$ so that

$$|q| = \exp(-\pi \text{Im}\tau) < 1 \tag{1.9}$$

[12, sections 21.1 and 21.6].

Each of the expansions (1.2)–(1.4) is valid in the strip

$$\left|\text{Im}\left(\frac{u}{K}\right)\right| < \text{Im}\left(\frac{iK'}{K}\right) = \text{Im}\tau \tag{1.10}$$

of the u-plane, and each of the expansions (1.5)–(1.7) is valid in the entire finite u-plane, with the poles of snu, cnu, or dnu, respectively, deleted [12, section 22.6], [3, p. 260], [4, sections 8.7 and 8.8].

In this paper, we obtain the following series expansions of $sn(u,k)$, $cn(u,k)$, and $dn(u,k)$:

$$\begin{aligned} sn(u,k) = {} & \frac{\pi}{2kK} \sum_{j=-m+1}^{m} \text{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] \\ & + \frac{2\pi q^{m+\frac{1}{2}}}{kK} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1-q^{2n+1}} \sin \frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{1.11}$$

$$\begin{aligned} cn(u,k) = {} & \frac{\pi i}{2kK} \sum_{j=-m+1}^{m} (-1)^j \text{cosec}\ \frac{\pi}{2K}[u-(2j-1)iK'] \\ & + (-1)^m \frac{2\pi q^{m+\frac{1}{2}}}{kK} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1+q^{2n+1}} \cos \frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{1.12}$$

$$\begin{aligned} dn(u,k) = {} & \frac{\pi}{2K} + \frac{\pi i}{2K} \sum_{j=-m+1}^{m} (-1)^j \left\{\cot \frac{\pi}{2K}[u-(2j-1)iK'] - i\alpha(j)\right\} \\ & + (-1)^m \frac{2\pi q^{2m+1}}{K} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1+q^{2n+2}} \cos \frac{(n+1)\pi u}{K}, \end{aligned} \tag{1.13}$$

where $m = 1, 2, 3, \ldots$, and

$$\alpha(j) = \begin{cases} -1 & \text{if } j \le 0, \\ +1 & \text{if } j > 0. \end{cases} \tag{1.14}$$

Each of the expansions (1.11)–(1.13) is valid (and can be used) in the strip

$$\begin{aligned} \left|\text{Im}\left(\frac{u}{K}\right)\right| & < \text{Im}\left[\frac{(2m+1)iK'}{K}\right] \\ & = \text{Im}[(2m+1)\tau], \end{aligned} \tag{1.15}$$

with the poles of snu, cnu, or dnu, respectively, located within that strip deleted. Strip (1.15) is $(2m+1)$ times wider than strip (1.10) where expansions (1.2)–(1.4) are valid, but, of course, it is "more narrow" than the entire finite u-plane where (with the exception of the poles of snu, cnu, or dnu, respectively) expansions (1.5)–(1.7) are valid.

Expansions (1.11)–(1.13) (obtained in this paper) can be viewed as intermediates between the known expansions (1.2)–(1.4) (on one hand) and expansions (1.5)–(1.7) (on the other hand), and they provide a link between these two sets of expansions known in the literature.

Because of the factors q^{2mn}, the infinite series expansions (1.11)–(1.13) should converge in strip (1.10) more rapidly (and more rapidly the larger m is) than the infinite series of (1.2)–(1.4), respectively. At the same time, the structure of the general terms of the infinite series of (1.11)–(1.13) is as simple as the structure of the general terms of the series (1.2)–(1.4).

Derivations and further discussion of expansions (1.11)–(1.13) follow.

2. Derivation of expansions (1.11) and (1.12). For each of the expansions (1.11)–(1.13), two different methods of obtaining each expansion can be used.

Namely, with regard to expansion (1.11) (for the sake of definiteness), one way to obtain it is to start from expansion (1.5) of $sn(u,k)$, to break the infinite sum of (1.5) into two parts,

$$\sum_{j=-m+1}^{m} \quad \text{and} \quad \left(\sum_{j=-\infty}^{-m} + \sum_{j=m+1}^{\infty}\right), \tag{2.1}$$

and then to represent the second part as the sine series in strip (1.15).

Another way to find (1.11) is to start from expansion (1.2) of $sn(u,k)$, to derive (1.11) for u in strip (1.10) first, and then to use analytic continuation to prove that (1.11) is indeed valid in strip (1.15).

In this paper, we shall use the first method to obtain (1.11) and (1.12) and the second method to obtain (1.13).

Therefore, let us write (1.5) as

$$sn(u,k) = \frac{\pi}{2kK}\sum_{j=-m+1}^{m} \operatorname{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] + R_s, \tag{2.2}$$

where

$$R_s = \frac{\pi}{2kK}\left(\sum_{j=-\infty}^{-m} + \sum_{j=m+1}^{\infty}\right) \operatorname{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] \tag{2.3}$$

and $m = 1, 2, 3, \ldots$.

The expression for R_s may be written as

$$R_s = \frac{\pi}{2kK}\sum_{j=m+1}^{\infty} f_{sn}^{(j)}(v) \quad \left(v = \frac{u}{K}\right), \tag{2.4}$$

where

$$f_{sn}^{(j)}(v) = \operatorname{cosec} \frac{\pi}{2}[v-(2j-1)\tau] + \operatorname{cosec} \frac{\pi}{2}[v+(2j-1)\tau] \tag{2.5}$$

$$= 4\cos\frac{(2j-1)\pi\tau}{2}\,\frac{\sin\left(\frac{\pi v}{2}\right)}{\cos(2j-1)\pi\tau-\cos\pi v} \tag{2.6}$$

$$= \frac{2(1+q^{2j-1})}{q^{j-\frac{1}{2}}}\,\frac{\sin\left(\frac{\pi v}{2}\right)}{\cos(2j-1)\pi\tau-\cos\pi v} \tag{2.7}$$

$$= \frac{4q^{j-\frac{1}{2}}(1+q^{2j-1})\sin\left(\frac{\pi v}{2}\right)}{1-2q^{2j-1}\cos\pi v+q^{4j-2}}. \tag{2.8}$$

The following expansion is cited (and used) in [3, p. 261]:

$$\frac{(1+a)\sin w}{1-2a\cos 2w+a^2} = \sum_{n=0}^{\infty} a^n \sin(2n+1)w. \tag{2.9}$$

Conditions on a and w for (2.9) to be valid are not indicated in [3], however. Let a and w in (2.9) be complex. By D'Alembert's test, one sees that the infinite series on the right-hand side of (2.9) is absolutely convergent if

$$|a|e^{2|\mathrm{Im} w|} < 1, \tag{2.10}$$

and it is divergent if

$$|a|e^{2|\mathrm{Im} w|} > 1. \tag{2.11}$$

Expansion (2.9) is valid if condition (2.10) is satisfied (and it is not valid if (2.11) is satisfied).

Using (2.9)–(2.11), we find that

$$\frac{(1+q^{2j-1})\sin\left(\frac{\pi v}{2}\right)}{1-2q^{2j-1}\cos\pi v+q^{4j-2}} = \sum_{n=0}^{\infty} q^{(2j-1)n}\sin\frac{(2n+1)\pi v}{2}, \tag{2.12}$$

and this is true if

$$|\mathrm{Im} v| < (2j-1)\ \mathrm{Im}\tau \tag{2.13}$$

and false if

$$|\mathrm{Im} v| > (2j-1)\ \mathrm{Im}\tau. \tag{2.14}$$

From (2.4)–(2.8) and (2.12), we get

$$R_s = \frac{2\pi q^{m+\frac{1}{2}}}{kK}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1-q^{2n+1}}\sin\frac{(2n+1)\pi u}{2K} \tag{2.15}$$

$(m = 1, 2, 3, \ldots)$ for

$$|\mathrm{Im} v| < (2m+1)\ \mathrm{Im}\tau. \tag{2.16}$$

In the process of obtaining (2.15), we changed the order of summation with respect to j and n in infinite sums. The validity of that change can be proved by the usual methods (see [1, section 8.23]).

From (2.2) and (2.15), expansion (1.11) immediately follows.

The derivation of expansion (1.12) for $cn(u,k)$ is similar, only instead of R_s we use

$$R_c = \frac{\pi i}{2kK}\left(\sum_{j=-\infty}^{-m} + \sum_{j=m+1}^{\infty}\right)(-1)^j \operatorname{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] \tag{2.17}$$

$$= \frac{\pi i}{2kK}\sum_{j=m+1}^{\infty}(-1)^j f_{cn}^{(j)}(v) \tag{2.18}$$

($m = 1, 2, 3, \ldots$), where

$$f_{cn}^{(j)}(v) = \operatorname{cosec} \frac{\pi}{2}[v-(2j-1)\tau] - \operatorname{cosec} \frac{\pi}{2}[v+(2j-1)\tau] \tag{2.19}$$

$$= 4\sin\frac{(2j-1)\pi\tau}{2}\,\frac{\cos\left(\frac{\pi v}{2}\right)}{\cos(2j-1)\pi\tau - \cos\pi v} \tag{2.20}$$

$$= \frac{2i(1-q^{2j-1})}{q^{j-\frac{1}{2}}}\,\frac{\cos\left(\frac{\pi v}{2}\right)}{\cos(2j-1)\pi\tau - \cos\pi v} \tag{2.21}$$

$$= \frac{4iq^{j-\frac{1}{2}}(1-q^{2j-1})\cos\left(\frac{\pi v}{2}\right)}{1-2q^{2j-1}\cos\pi v + q^{4j-2}}, \tag{2.22}$$

and instead of (2.9) we use the expansion

$$\frac{(1-a)\cos w}{1-2a\cos 2w + a^2} = \sum_{n=0}^{\infty} a^n \cos(2n+1)w \tag{2.23}$$

(which one can easily obtain from (2.9) by replacing a by $-a$ and w by $w + (\pi/2)$, respectively).

Continuing in this way, we find that

$$R_c = (-1)^m \frac{2\pi q^{m+\frac{1}{2}}}{kK}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K} \tag{2.24}$$

and finally obtain expansion (1.12) for u in strip (1.15).

In addition to R_s and R_c, below we shall also use (in section 4) the notation

$$R_d = (-1)^m \frac{2\pi q^{2m+1}}{K}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K} \tag{2.25}$$

($m = 1, 2, 3, \ldots$) for the term with the infinite series in n of the right-hand side of (1.13).

3. Derivation of expansion (1.13). As mentioned in the previous section, here we shall use a different approach. Namely, we start from expansion (1.4) and write it as

$$dn(u,k) = \frac{\pi}{2K} + \frac{2\pi}{K}\sum_{n=0}^{\infty}\frac{q^{n+1}}{1+q^{2n+2}}\cos(n+1)\pi v \tag{3.1}$$
$$+ F_{dn}^{(m)}(v) - F_{dn}^{(m)}(v),$$

with

$$F_{dn}^{(m)}(v) = \frac{\pi i}{2K} \sum_{j=-m+1}^{m} (-1)^j \cot \frac{\pi}{2}[v - (2j-1)\tau] \tag{3.2}$$

$(m = 1, 2, 3, \ldots)$.

The expression for $F_{dn}^{(m)}$ can be rewritten as

$$F_{dn}^{(m)}(v) = \frac{\pi i}{2K} \sum_{j=1}^{m} (-1)^j f_{dn}^{(j)}(v), \tag{3.3}$$

where

$$\begin{aligned} f_{dn}^{(j)}(v) &= \cot \frac{\pi}{2}[v - (2j-1)\tau] - \cot \frac{\pi}{2}[v + (2j-1)\tau] && (3.4)\\ &= 2\sin[(2j-1)\pi\tau] \frac{1}{\cos(2j-1)\pi\tau - \cos \pi v} && (3.5)\\ &= \frac{i(1-q^{4j-2})}{q^{2j-1}} \frac{1}{\cos(2j-1)\pi\tau - \cos \pi v} && (3.6)\\ &= \frac{2i(1-q^{4j-2})}{1 - 2q^{2j-1}\cos \pi v + q^{4j-2}}. && (3.7)\end{aligned}$$

Multiplying (2.9) by $(1-a)\sin w$ and (2.23) by $(1+a)\cos w$ and adding the results, we obtain

$$\frac{(1-a^2)}{1 - 2a\cos 2w + a^2} = 1 + 2\sum_{n=1}^{\infty} a^n \cos 2nw, \tag{3.8}$$

which is valid under condition (2.10) (and is not valid if (2.11) is satisfied).

From (3.8), it follows that

$$\frac{1 - q^{4j-2}}{1 - 2q^{2j-1}\cos \pi v + q^{4j-2}} = 1 + 2\sum_{n=1}^{\infty} q^{(2j-1)n} \cos n\pi v \tag{3.9}$$

$(j = 1, 2, 3, \ldots)$, and this is true under condition (2.13) (and not true if (2.14) is satisfied).

From (3.3), (3.4)–(3.7), and (3.9), we get

$$\frac{K}{\pi} F_{dn}^{(m)}(v) = E(m, n = 0) + 2\sum_{n=1}^{\infty} E(m,n) \cos n\pi v \tag{3.10}$$

$(m = 1, 2, 3, \ldots)$, where

$$E(m,n) = \sum_{j=1}^{m} (-1)^{j-1} q^{(2j-1)n} \tag{3.11}$$

$(n = 0, 1, 2, \ldots)$, and the result (3.10) is certainly valid for $v(= u/K)$ in strip (1.10).

From (3.11), we obtain

$$E(m,n) = q^n \frac{1 - (-1)^m q^{2nm}}{1 + q^{2n}} \tag{3.12}$$

$(m = 1, 2, 3, \ldots, n = 0, 1, 2, \ldots)$. In particular,

$$E(m, n = 0) = \frac{1 - (-1)^m}{2} \tag{3.13}$$

$(m = 1, 2, 3, \ldots)$.

Equation (3.10) can now be rewritten as

$$F_{dn}^{(m)}(v) = \frac{\pi}{2K}[1 - (-1)^m] + \frac{2\pi}{K}\sum_{n=0}^{\infty} q^{n+1} \cdot \frac{1 - (-1)^m q^{(2n+2)m}}{1 + q^{2n+2}} \cos(n+1)\pi v \tag{3.14}$$

$(m = 1, 2, 3, \ldots, v$ in strip (1.10)).

We now substitute for the first $F_{dn}^{(m)}$ of (3.1) its expression (3.2) and for the second $F_{dn}^{(m)}$ its expression (3.14) to get

$$dn(u, k) = (-1)^m \frac{\pi}{2K} + \frac{\pi i}{2K} \sum_{j=-m+1}^{m} (-1)^j \cdot \cot \frac{\pi}{2K}[u - (2j-1)iK'] + (-1)^m \frac{2\pi q^{2m+1}}{K} \cdot \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1 + q^{2n+2}} \cos \frac{(n+1)\pi u}{K} \tag{3.15}$$

$(m = 1, 2, 3, \ldots)$ for u in strip (1.10).

Taking into account the fact that

$$(-1)^m = 1 + 2\sum_{j=1}^{m}(-1)^j = \sum_{j=-m+1}^{m} (-1)^j \alpha(j) + 1 \tag{3.16}$$

$(m = 1, 2, 3, \ldots)$, one can now rewrite (3.15) in the form of (1.13) for u in strip (1.10). Finally, to prove that (1.13) is valid in strip (1.15), one can use analytic continuation.

4. Estimates for *n*-series in (1.11)–(1.13). In (1.11), the n-series is R_s (see (1.11) and (2.15)). For u in strip (1.15), we can write

$$\operatorname{Im}\left(\frac{u}{K}\right) = \gamma \operatorname{Im}[(2m+1)\tau], \tag{4.1}$$

where $-1 < \gamma < +1$ and $m = 1, 2, 3, \ldots$. For u in this strip, we obtain

$$\left|\sin \frac{(2n+1)\pi u}{2K}\right| \le |q|^{-(2m+1)(2n+1)|\gamma|/2} \tag{4.2}$$

and then

$$|R_s| \le \frac{2\pi}{|kK|} \frac{|q|^{(m+\frac{1}{2})(1-|\gamma|)}}{1 - |q|} \frac{1}{1 - |q|^{(2m+1)(1-|\gamma|)}} \tag{4.3}$$

$(m = 1, 2, 3, \ldots)$.

Similarly, for the n-series of (1.12), which is R_c (see (1.12) and (2.24)), we find the same estimate (4.3) in strip (1.15).

Also, in strip (1.15),

$$\left|\cos\frac{(n+1)\pi u}{K}\right| \leq |q|^{-(2m+1)(n+1)|\gamma|}, \tag{4.4}$$

from which we obtain the estimate for R_d (see (2.25)):

$$|R_d| \leq \frac{2\pi}{|K|}\frac{|q|^{(2m+1)(1-|\gamma|)}}{1-|q|^2}\frac{1}{1-|q|^{(2m+1)(1-|\gamma|)}} \tag{4.5}$$

$(m = 1, 2, 3, \ldots)$ in strip (1.15).

If instead of (4.2), we use the estimate

$$\left|\sin\frac{(2n+1)\pi u}{2K}\right| \leq \frac{1}{2}[|q|^{-(2m+1)(2n+1)\gamma/2} + |q|^{(2m+1)(2n+1)\gamma/2}], \tag{4.6}$$

then instead of (4.3), the following bound for $|R_s|$ is obtained:

$$\begin{aligned}|R_s| \leq{}& \frac{\pi}{|kK|}\frac{|q|^{(m+\frac{1}{2})(1-|\gamma|)}}{1-|q|}\\ &\cdot\frac{1}{1-|q|^{(2m+1)(1-|\gamma|)}}\left\{1+|q|^{(2m+1)|\gamma|}\frac{1-|q|^{(2m+1)(1-|\gamma|)}}{1-|q|^{(2m+1)(1+|\gamma|)}}\right\},\end{aligned} \tag{4.7}$$

and this bound is true for $|R_c|$ as well.

Also, if instead of (4.4), we use the estimate

$$\left|\cos\frac{(n+1)\pi u}{K}\right| \leq \frac{1}{2}[|q|^{-(2m+1)(n+1)\gamma} + |q|^{(2m+1)(n+1)\gamma}], \tag{4.8}$$

then instead of (4.5), the following bound for $|R_d|$ is found:

$$\begin{aligned}|R_d| \leq{}& \frac{\pi}{|K|}\frac{|q|^{(2m+1)(1-|\gamma|)}}{1-|q|^2}\\ &\cdot\frac{1}{1-|q|^{(2m+1)(1-|\gamma|)}}\left\{1+|q|^{(4m+2)|\gamma|}\frac{1-|q|^{(2m+1)(1-|\gamma|)}}{1-|q|^{(2m+1)(1+|\gamma|)}}\right\}.\end{aligned} \tag{4.9}$$

5. Obtaining series expansions (1.5)–(1.7) from equations (1.11)–(1.13). From (4.3), it follows that $|R_s| \to 0$ as $m \to \infty$. Accordingly, the limiting form of (1.11) as $m \to \infty$ is (1.5) with the sum written as

$$\lim_{m\to\infty}\sum_{j=-m+1}^{m} \tag{5.1}$$

(and also, of course, (1.5) is known to be true with the sum written as it is written in (1.5)).

Similarly, (1.6) with the sum in it written as in (5.1) is the limiting form of (1.12) as $m \to \infty$.

The bound (4.3) (or the bound (4.7)) for $|R_s|$ and $|R_c|$ can be looked upon as an upper bound for the absolute values of the remainders of the approximations

$$sn(u,k) \simeq \frac{\pi}{2kK} \sum_{j=-m+1}^{m} \operatorname{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] \tag{5.2}$$

and

$$cn(u,k) \simeq \frac{\pi i}{2kK} \sum_{j=-m+1}^{m} (-1)^j \operatorname{cosec} \frac{\pi}{2K}[u-(2j-1)iK'] \tag{5.3}$$

in strip (1.15); these approximations can be viewed as being obtained by truncating the infinite series of (1.5) and (1.6) (in the manner indicated by the extreme values of j in the sums of (5.2) and (5.3)).

Now taking into account (4.5), we find the following expansion as the limiting form of (1.13) as $m \to \infty$:

$$dn(u,k) = \frac{\pi}{2K} + \frac{\pi i}{2K} \lim_{m\to\infty} \sum_{j=-m+1}^{m} (-1)^j \left\{ \cot \frac{\pi}{2K}[u-(2j-1)iK'] - i\alpha(j) \right\} \tag{5.4}$$

($|u| < \infty$, poles of dnu deleted).

This expansion looks different from (1.7) but is equivalent to it. Indeed, for any fixed u of (5.4),

$$\lim_{m\to+\infty} \cot \frac{\pi}{2K}[u+(2m+1)\pi iK'] = -i, \tag{5.5}$$

and by (5.5), equation (5.4) can be reduced to (1.7).

6. Other forms of expansions (1.11)–(1.13). Here are alternative forms of expansions (1.11)–(1.13):

$$\begin{aligned} sn(u,k) = {} & \frac{2\pi q^{1/2}}{kK} \sum_{j=1}^{m} \frac{q^{j-1}(1+q^{2j-1}) \sin \dfrac{\pi u}{2K}}{1-2q^{2j-1}\cos\left(\dfrac{\pi u}{K}\right)+q^{4j-2}} \\ & + \frac{2\pi q^{m+\frac12}}{kK} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1-q^{2n+1}} \sin \frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{6.1}$$

$$\begin{aligned} sn(u,k) = {} & \frac{\pi q^{1/2}}{kK} \sum_{j=1}^{m} \frac{1+q^{2j-1}}{q^j} \frac{\sin \frac{\pi u}{2K}}{\cosh\left[\dfrac{(2j-1)\pi K'}{K}\right] - \cos\left(\dfrac{\pi u}{K}\right)} \\ & + \frac{2\pi}{kK} q^{m+\frac12} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1-q^{2n+1}} \sin \frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{6.2}$$

$$\begin{aligned} cn(u,k) = {} & \frac{2\pi q^{1/2}}{kK} \sum_{j=1}^{m} (-1)^{j-1} \frac{q^{j-1}(1-q^{2j-1}) \cos \dfrac{\pi u}{2K}}{1-2q^{2j-1}\cos\left(\dfrac{\pi u}{K}\right)+q^{4j-2}} \\ & + (-1)^m \frac{2\pi q^{m+\frac12}}{kK} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1+q^{2n+1}} \cos \frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{6.3}$$

$$(6.4)\quad cn(u,k) = \frac{\pi q^{1/2}}{kK}\sum_{j=1}^{m}(-1)^{j-1}\frac{1-q^{2j-1}}{q^j}\frac{\cos\dfrac{\pi u}{2K}}{\cosh\left[\dfrac{(2j-1)\pi K'}{K}\right]-\cos\left(\dfrac{\pi u}{K}\right)}$$
$$+(-1)^m\frac{2\pi q^{m+\frac{1}{2}}}{kK}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K},$$

$$(6.5)\quad dn(u,k) = (-1)^m\frac{\pi}{2K}+\frac{\pi}{K}\sum_{j=1}^{m}(-1)^{j-1}\frac{1-q^{4j-2}}{1-2q^{2j-1}\cos\left(\dfrac{\pi u}{K}\right)+q^{4j-2}}$$
$$+(-1)^m\frac{2\pi}{K}q^{2m+1}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K},$$

$$(6.6)\quad dn(u,k) = (-1)^m\frac{\pi}{2K}+\frac{\pi}{2K}\sum_{j=1}^{m}(-1)^{j-1}\frac{1-q^{4j-2}}{q^{2j-1}}\frac{1}{\cosh\left[\dfrac{(2j-1)\pi K'}{K}\right]-\cos\dfrac{\pi u}{K}}$$
$$+(-1)^m\frac{2\pi q^{2m+1}}{K}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K},$$

$$(6.7)\quad dn(u,k) = \frac{\pi}{2K}+\frac{2\pi}{K}\sum_{j=1}^{m}(-1)^{j-1}q^{2j-1}\frac{\cos\dfrac{\pi u}{K}-q^{2j-1}}{1-2q^{2j-1}\cos\dfrac{\pi u}{K}+q^{4j-2}}$$
$$+(-1)^m\frac{2\pi}{K}q^{2m+1}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K},$$

and

$$(6.8)\qquad dn(u,k) = \frac{\pi}{2K}+\frac{\pi}{K}\sum_{j=1}^{m}(-1)^{j-1}\frac{\cos\dfrac{\pi u}{K}-q^{2j-1}}{\cosh\dfrac{(2j-1)\pi K'}{K}-\cos\dfrac{\pi u}{K}}$$
$$+(-1)^m\frac{2\pi q^{2m+1}}{K}\sum_{n=0}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K}.$$

One or another expansion for the same function (sn, cn, or dn) may be preferable to use in certain cases. For example, if τ is purely imaginary (the usual case in applications of elliptic functions in physics and engineering [5]), then K and K' are real, and expansions (6.1) or (6.2) may be more convenient to use than (1.11) for u real.

At $m=\infty$, expansions (6.1)–(6.4) and (6.7) and (6.8) yield

$$(6.9)\qquad sn(u,k) = \frac{2\pi q^{1/2}}{kK}\sum_{j=1}^{\infty}\frac{q^{j-1}(1+q^{2j-1})\sin\dfrac{\pi u}{2K}}{1-2q^{2j-1}\cos\left(\dfrac{\pi u}{K}\right)+q^{4j-2}},$$

$$(6.10)\qquad sn(u,k) = \frac{\pi q^{1/2}}{kK}\sum_{j=1}^{\infty}\frac{1+q^{2j-1}}{q^j}\frac{\sin\dfrac{\pi u}{2K}}{\cosh\left[\dfrac{(2j-1)\pi K'}{K}\right]-\cos\left(\dfrac{\pi u}{K}\right)},$$

$$cn(u,k) = \frac{2\pi q^{1/2}}{kK}\sum_{j=1}^{\infty}(-1)^{j-1}\frac{q^{j-1}(1-q^{2j-1})\cos\dfrac{\pi u}{2K}}{1-2q^{2j-1}\cos\left(\dfrac{\pi u}{K}\right)+q^{4j-2}}, \tag{6.11}$$

$$cn(u,k) = \frac{\pi q^{1/2}}{kK}\sum_{j=1}^{\infty}(-1)^{j-1}\frac{1-q^{2j-1}}{q^j}\frac{\cos\dfrac{\pi u}{2K}}{\cosh\left[\dfrac{(2j-1)\pi K'}{K}\right]-\cos\left(\dfrac{\pi u}{K}\right)}, \tag{6.12}$$

$$dn(u,k) = \frac{\pi}{2K}+\frac{2\pi}{K}\sum_{j=1}^{\infty}(-1)^{j-1}q^{2j-1}\frac{\cos\dfrac{\pi u}{K}-q^{2j-1}}{1-2q^{2j-1}\cos\dfrac{\pi u}{K}+q^{4j-2}}, \tag{6.13}$$

and

$$dn(u,k) = \frac{\pi}{2K}+\frac{\pi}{K}\sum_{j=1}^{\infty}(-1)^{j-1}\frac{\cos\dfrac{\pi u}{K}-q^{2j-1}}{\cosh\dfrac{(2j-1)\pi K'}{K}-\cos\dfrac{\pi u}{K}}, \tag{6.14}$$

respectively ($|u| < \infty$, poles of $sn(u,k)$, $cn(u,k)$, and $dn(u,k)$, respectively, are deleted).

7. Sums of some infinite series. By putting such specific values of u as $u = 0$, $K/2, K, \ldots$ or $u = iK'/2, iK', \ldots$ into the expansions derived above we can obtain results for sums of interesting infinite series. As examples, we cite here the following two results:

$$\sum_{n=0}^{\infty}(-1)^{E(n/2)}\frac{q^{n+m+2nm}}{1-q^{2n+1}} = \frac{kK}{\pi[2q(1+k')]^{1/2}}-\sum_{j=1}^{m}\frac{q^{j-1}(1+q^{2j-1})}{1+q^{4j-2}}, \tag{7.1}$$

$$\sum_{n=1}^{\infty}\frac{q^{(2m+1)n}}{1+q^{2n}} = (-1)^m\left(\frac{K}{2\pi}-\frac{1}{4}\right)-(-1)^m\sum_{j=1}^{m}(-1)^{j-1}\frac{q^{2j-1}}{1-q^{2j-1}}, \tag{7.2}$$

where $m = 1, 2, 3, \ldots$ and $E(x)$ denotes the integral part of the real number x.

This results (7.1) and (7.2) are obtained if one puts $u = K/2$ in (6.1) and $u = 0$ in (6.5), respectively.

8. On the use of trigonometric series expansions of *snu*, *cnu*, and *dnu* in evaluating integrals. In applications, the importance of the trigonometric series expansions of *snu*, *cnu*, and *dnu* is that they can often be used for evaluating integrals of these functions multiplied by another function, like the integral

$$J = \int_0^K \sin\frac{\pi u}{K}\, sn(u,k)\, du \tag{8.1}$$

or similar integrals. Namely, by using trigonometric series expansions of *snu*, *cnu*, and *dnu*, one can often represent the integrals in question as infinite series, or asymptotic expressions for these integrals (say, corresponding to k small) can be found.

Using the example of the integral in (8.1), we shall now explain how the use of different expansions of $sn(u,k)$ may be advantageous. Namely, if we use the expansion (1.2) in (8.1), we get

$$J = 8q^{1/2}k^{-1}\sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{(2n-1)(2n+3)}\frac{q^n}{1-q^{2n+1}}. \tag{8.2}$$

If, on the other hand, we use the expansion (6.2) of $sn(u,k)$ in (8.1), then we get

$$J=\sum_{j=1}^{m} J_j + 8q^{m+\frac{1}{2}}k^{-1}\sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{(2n-1)(2n+3)}\frac{q^{(2m+1)n}}{1-q^{2n+1}}, \tag{8.3}$$

where the sum in j (from $j=1$ to $j=m$) is the result of integration of the term with the sum in j of (6.2) (multiplied by $\sin(\pi u/K)$).

The concrete expressions for J_j are of no importance for the sake of our discussion, and therefore they are not presented here. What is important is that because of the presence of the factors $q^{(2m+1)n}$, the infinite series (in n) of (8.3) is more rapidly convergent than the infinite series of (8.2) (even at $m=1$). Thus using expansion (6.2) for $sn(u,k)$ instead of (1.2) in (8.1), one can obtain a more rapidly convergent series for J.

Of course, if one has already obtained an infinite series representation for some integral, it is possible to try to increase the rapidity of convergence of that series. However, in applications one often wants to (or can) obtain not an entire infinite series expansion but only an asymptotic formula expression (say, corresponding to small k) for an integral. In such cases, the use of a more rapidly convergent series expansion for snu, cnu, and dnu (like (1.11)–(1.13) and (6.1)–(6.8)) is very expedient. Also, the simple structure of the series in n of (1.11)–(1.13) and (6.1)–(6.8) allows for convenient evaluation of the remainder (of the asymptotic formula obtained).

In more complicated situations, if one needs to solve an integral equation with its kernel expressed in terms of sn, cn, or dn, the use of the series expansions (1.11)–(1.13) and (6.1)–(6.8) with their rapidly convergent series in n can be of advantage as well.

9. Expansions of *cdu*, *sdu*, ..., *scu* similar to expansions of *snu*, *cnu*, and *dnu* (1.11)–(1.13). They are as follows:

$$\begin{aligned} cd(u,k) &= \frac{\pi}{2kK}\sum_{j=-m+1}^{m}\sec\frac{\pi}{2K}[u-(2j-1)iK'] \\ &\quad + \frac{2\pi q^{m+\frac{1}{2}}}{kK}\sum_{n=0}^{\infty}(-1)^n\frac{q^{(2m+1)n}}{1-q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{9.1}$$

$$\begin{aligned} sd(u,k) &= -\frac{\pi i}{2kk'K}\sum_{j=-m+1}^{m}(-1)^j\sec\frac{\pi}{2K}[u-(2j-1)iK'] \\ &\quad + (-1)^m\frac{2\pi q^{m+\frac{1}{2}}}{kk'K}\sum_{n=0}^{\infty}(-1)^n\frac{q^{(2m+1)n}}{1+q^{2n+1}}\sin\frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{9.2}$$

$$\begin{aligned} nd(u,k) &= \frac{\pi}{2k'K} + \frac{\pi i}{2k'K}\sum_{j=-m+1}^{m}(-1)^{j+1}\left\{\tan\frac{\pi}{2K}[u-(2j-1)iK'] + i\alpha(j)\right\} \\ &\quad + (-1)^m\frac{2\pi q^{2m+1}}{k'K}\sum_{n=0}^{\infty}(-1)^{n+1}\frac{q^{(2m+1)n}}{1+q^{2n+2}}\cos\frac{(n+1)\pi u}{K}, \end{aligned} \tag{9.3}$$

$$\begin{aligned} ns(u,k) &= \frac{\pi}{2K}\sum_{j=-m}^{m}\operatorname{cosec}\frac{\pi}{2K}(u-2jiK') \\ &\quad + \frac{2\pi}{K}\sum_{n=0}^{\infty}\frac{q^{(2n+1)(m+1)}}{1-q^{2n+1}}\sin\frac{(2n+1)\pi u}{2K}, \end{aligned} \tag{9.4}$$

$$(9.5)\quad dc(u,k) = \frac{\pi}{2K}\sum_{j=-m}^{m} \sec\frac{\pi}{2K}(u-2jiK') + \frac{2\pi}{K}\sum_{n=0}^{\infty}(-1)^n\frac{q^{(2n+1)(m+1)}}{1-q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K},$$

$$(9.6)\quad ds(u,k) = \frac{\pi}{2K}\sum_{j=-m}^{m}(-1)^j\text{cosec }\frac{\pi}{2K}(u-2jiK') + (-1)^{m+1}\frac{2\pi}{K}\sum_{n=0}^{\infty}\frac{q^{(2n+1)(m+1)}}{1+q^{2n+1}}\sin\frac{(2n+1)\pi u}{2K},$$

$$(9.7)\quad nc(u,k) = \frac{\pi}{2k'K}\sum_{j=-m}^{m}(-1)^j\sec\frac{\pi}{2K}(u-2jiK') + (-1)^{m+1}\frac{2\pi}{k'K}\sum_{n=0}^{\infty}(-1)^n\frac{q^{(2n+1)(m+1)}}{1+q^{2n+1}}\cos\frac{(2n+1)\pi u}{2K},$$

$$(9.8)\quad cs(u,k) = \frac{\pi}{2K}\sum_{j=-m}^{m}(-1)^j\cot\frac{\pi}{2K}(u-2jiK') + (-1)^{m+1}\frac{2\pi}{K}\sum_{n=1}^{\infty}\frac{q^{2n(m+1)}}{1+q^{2n}}\sin\frac{n\pi u}{K},$$

$$(9.9)\quad sc(u,k) = \frac{\pi}{2k'K}\sum_{j=-m}^{m}(-1)^j\tan\frac{\pi}{2K}(u-2jiK') + (-1)^{m}\frac{2\pi}{k'K}\sum_{n=1}^{\infty}(-1)^n\frac{q^{2n(m+1)}}{1+q^{2n}}\sin\frac{n\pi u}{K}.$$

In (9.1)–(9.3), $m = 1, 2, 3, \ldots$, and these expansions are valid in strip (1.15) (with $m = 1, 2, 3, \ldots$). In (9.4)–(9.9), $m = 0, 1, 2, \ldots$, and these are valid in the strip

$$(9.10)\qquad \left|\text{Im}\left(\frac{u}{K}\right)\right| < \text{Im}[(2m+2)\tau]$$

(with $m = 0, 1, 2, \ldots$).

Equations (9.1), (9.2), and (9.3) are obtained by replacing u by $u+K$ in (1.11), (1.12), and (1.13), respectively. Equations (9.4), (9.6), and (9.8) are obtained from the trigonometric series expansions of $ns(u,k)$, $ds(u,k)$, and $cs(u,k)$, respectively, given in [12, section 22.61] by a method similar to how (1.13) was obtained from (1.4) in section 3 above. Instead of $F_{dn}^{(m)}(v)$ (see equation (3.2) above), however, we use

$$(9.11)\qquad F_{ns}^{(m)}(v) = \frac{\pi}{2K}\sum_{j=1}^{m}\left[\text{cosec }\frac{\pi}{2}(v-2j\tau) + \text{cosec }\frac{\pi}{2}(v+2j\tau)\right],$$

$$(9.12)\qquad F_{ds}^{(m)}(v) = \frac{\pi}{2K}\sum_{j=1}^{m}(-1)^j\left[\text{cosec }\frac{\pi}{2}(v-2j\tau) + \text{cosec }\frac{\pi}{2}(v+2j\tau)\right],$$

and

$$(9.13)\qquad F_{cs}^{(m)}(v) = \frac{\pi}{2K}\sum_{j=1}^{m}(-1)^j\left[\cot\frac{\pi}{2}(v-2j\tau) + \cot\frac{\pi}{2}(v+2j\tau)\right]$$

($m = 1, 2, 3, \ldots$) to derive (9.4), (9.6), and (9.8), respectively.

Finally, one obtains equations (9.5), (9.7), and (9.9) by replacing u by $u + K$ in (9.4), (9.6), and (9.8), respectively.

10. Expansions in asymetric strips. Here we shall give pertinent explanations using the example of $sn(u, k)$.

In the $v = u/K$-plane, the function $sn(u, k)$ has poles at points

$$v = 2n + (2m + 1)\tau, \quad n = 0, \pm1, \pm2, \ldots, \quad m = 0, \pm1, \pm2, \ldots, \tag{10.1}$$

(see (1.1)), and all of these poles are located on the horizontal lines

$$\mathrm{Im}v = \mathrm{Im}[(2m + 1)\tau], \quad m = 0, \pm1, \pm2, \ldots, \tag{10.2}$$

in the v-plane.

Now expansion (1.2) (known in the literature) and expansion (1.11) of $sn(u, k)$ (derived in this paper) are valid in the strips

$$|\mathrm{Im}v| < \mathrm{Im}\tau \tag{10.3}$$

and

$$|\mathrm{Im}v| < \mathrm{Im}[(2m + 1)\tau], \quad m = 1, 2, 3, \ldots, \tag{10.4}$$

respectively, and each of these strips is seen to be symmetric with respect to the Rev-axis in the v-plane.

However, using the set of expansions derived in this paper, one can easily obtain expansions of $sn(u, k)$ valid in asymmetric (with respect to the Rev-axis in the v-plane) strips as well. Namely, all the strips between the (horizontal) lines (10.2) of poles of $sn(u, k)$ in the v-plane belong to one of the following categories:

$$\mathrm{Im}[(2s - 1)\tau] < \mathrm{Im}v < \mathrm{Im}[(2s + 1)\tau], \tag{10.5}$$

where $s = 0, \pm1, \pm2, \ldots$, or

$$\mathrm{Im}[(2s - 2m - 1)\tau] < \mathrm{Im}v < \mathrm{Im}[(2s + 2m + 1)\tau], \tag{10.6}$$

where $s = 0, \pm1, \pm2, \ldots$ and $m = 1, 2, 3, \ldots$, or

$$\mathrm{Im}[(2s - 2m + 1)\tau] < \mathrm{Im}v < \mathrm{Im}[(2s + 2m + 1)\tau], \tag{10.7}$$

where $s = 0, \pm1, \pm2, \ldots$ and $m = 1, 2, 3, \ldots$.

To obtain the expansion of $sn(u, k)$ valid in strip (10.5), we use [12, section 22.34]

$$sn(u, k) = sn(u - 2siK', k) \tag{10.8}$$

for $s = 0, \pm1, \pm2, \ldots$. Using (1.2), we find that

$$sn(u, k) = \frac{2\pi}{kK} \sum_{n=0}^{\infty} \frac{q^{n+\frac{1}{2}}}{1 - q^{2n+1}} \sin \frac{(2n + 1)\pi(u - 2siK')}{2K} \tag{10.9}$$

in strip (10.5). Similarly, using (1.11), we find that

$$\begin{aligned} sn(u, k) = \frac{\pi}{2kK} \sum_{j=-m+1}^{m} \operatorname{cosec} \frac{\pi}{2K}[u - (2j - 1)iK' - 2siK'] \\ + \frac{2\pi q^{m+\frac{1}{2}}}{kK} \sum_{n=0}^{\infty} \frac{q^{(2m+1)n}}{1 - q^{2n+1}} \sin \frac{(2n + 1)\pi(u - 2siK')}{2K} \end{aligned} \tag{10.10}$$

for v in strip (10.6).

Finally, using (see [2, equation 122.24])

$$sn(u,k) = k^{-1} dc[u - K - (2s+1)iK', k] \tag{10.11}$$

for $s = 0, \pm 1, \pm 2, \ldots$ and (9.5), we obtain

$$\begin{aligned} sn(u,k) = {} & \frac{\pi}{2kK} \sum_{j=-m+1}^{m-1} \operatorname{cosec} \frac{\pi}{2K}[u - 2jiK' - (2s+1)iK'] \\ & + \frac{2\pi}{kK} \sum_{n=0}^{\infty} \frac{q^{(2n+1)m}}{1 - q^{2n+1}} \sin \frac{(2n+1)\pi[u - (2s+1)iK']}{2K} \end{aligned} \tag{10.12}$$

in strip (10.7).

11. Expansions involving powers of q'. In practical applications of elliptic functions to problems of physics and engineering, τ is usually purely imaginary and $0 < q < 1$. Expansions involving powers of q (such as (1.2)–(1.4)) are used in applications only when $q \ll 1$. Otherwise, expansions involving powers of the parameter q', where [12, section 21.51]

$$q' = e^{\pi i \tau'}, \quad \tau' = -\frac{1}{\tau} = \frac{iK}{K'}, \tag{11.1}$$

are used: when $q \to 1, q' \to 0$.

Expansions involving powers of q' can be obtained (in the general case—not necessarily for purely imaginary τ) from the expansions involving powers of q by the use of Jacobi's imaginary transformation [12, section 22.4]. For example, we have [12, section 22.4]

$$sn(u,k) = -isc(iu, k'), \tag{11.2}$$

and using (9.9) in (11.2), we obtain

$$\begin{aligned} sn(u,k) = {} & \frac{\pi}{2kK'} \sum_{j=-m}^{m} (-1)^j \tanh \frac{\pi}{2K'}(u - 2jK) \\ & + (-1)^m \frac{2\pi}{kK'} \sum_{n=1}^{\infty} (-1)^n \frac{(q')^{2n(m+1)}}{1 + (q')^{2n}} \sinh \frac{n\pi u}{K'} \end{aligned} \tag{11.3}$$

for u in the strip

$$\left| \operatorname{Im}\left(\frac{iu}{K'}\right) \right| < \operatorname{Im}[(2m+2)\tau'] \tag{11.4}$$

with $m = 0, 1, 2, \ldots$, etc.

12. Concluding remarks and acknowledgments. In October 1993, this author submitted a paper entitled "New trigonometric expansions of Jacobian Elliptic Functions *snu*, *cnu*, *dnu*" to *SIAM J. Math. Anal.* That paper (which was the first version of this one) contained, in particular, results given by equations (1.11)–(1.15), (3.2), (3.3), (3.6), (3.7), (3.14) (written for $\xi = \operatorname{Re} v$ instead of v), (3.15), (5.4), (6.1)–(6.8), (7.1), and (7.2) of this paper, as well as a number of other results.

The paper was originally reviewed by one referee and then (in a revised version entitled "On trigonometric series expansions of Jacobian elliptic functions *snu*, *cnu*, *dnu*") by two referees. The referees made valuable comments, for which the author is grateful. The present paper is a revised and expanded version of this previous submission. The referees' comments were taken into account, some material from the first version was deleted, and new material was added. As a result, this paper is twice as long as the original submission.

This paper is based on simple ideas, which—as is often the case with simple ideas—work rather well and give good, usable results. Similar ideas worked out with respect to expansions for meromorphic functions and expansions for the solutions of the functional equations of the second kind (in particular, the Fredholm integral equation) are conveyed in this author's papers [8], [9], [10], and [11]. Also, in the abstract [6],[1] expansions for the logarithmic derivatives of the $\vartheta_{1-4}(\ldots)$ functions (a development similar to the one set forth in the present paper) are described.

The results (1.11)–(1.15), (5.4), and (6.2) of this paper are given (without derivation) and some other results of this paper are mentioned in the abstract [7].

The present paper was reviewed by a referee who has read it extremely thoroughly. The referee checked a large number of the formulas of the paper and also suggested numerous stylistic improvements. The author is very grateful to this referee.

REFERENCES

[1] T. M. Apostol, *Mathematical Analysis*, 2nd. ed., Addison–Wesley, Reading, MA, 1974.

[2] P. F. Byrd and M. D.Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, Springer-Verlag, Berlin, 1954.

[3] H. Hancock, *Lectures on the Theory of Elliptic Functions*, Vol. 1, John Wiley, New York, 1910.

[4] D. F. Lawden, *Elliptic Functions and Applications*, Springer-Verlag, New York, 1989.

[5] F. Oberhettinger and W. Magnus, *Anwendung der Elliptischen Funktionen in Physik und Technik*, Springer-Verlag, Berlin, 1949 (in German).

[6] D. S. Tselnik, *Expansions for the logarithmic derivatives of the theta$_{1-4}$ functions*, Abstracts Amer. Math. Soc., 14 (1993), p. 700.

[7] D. S. Tselnik, *New trigonometric expansions of Jacobian elliptic functions snu, cnu, dnu*, Abstracts Amer. Math. Soc., 15 (1994), p. 453.

[8] D. S. Tselnik, *Representation and series expansions for meromorphic functions*, Complex Variables Theory Appl., 25 (1994), pp. 159–171.

[9] D. S. Tselnik, *A simple bound for the remainder of the Neumann series in the case of a self-adjoint compact operator*, Appl. Math. Lett., 7 (1994), pp. 71–74.

[10] D. S. Tselnik, *A bound for the remainder of the Hilbert–Schmidt series and other results on representation of solutions to the functional equation of the second kind with a self-adjoint compact operator as an infinite series*, Comput. Math. Appl., 29 (1995), pp. 61–68.

[11] D. S. Tselnik, *On series expansions for meromorphic functions*, J. Math. Anal. Appl., 193 (1995), pp. 522–542.

[12] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., University Press, Cambridge, UK, 1965.

[1] *Note*: In the first line of the text of [6], please read $\vartheta'_j(z,q)/\vartheta_j(z,q)$ and $j = 1 - 4$ instead of $\vartheta'_i(z,q)/\vartheta_i(z,q)$ and $i = 1 - 4$, respectively.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 731–748, May 1997

013

SMOOTH REFINABLE FUNCTIONS PROVIDE GOOD APPROXIMATION ORDERS*

AMOS RON[†]

Abstract. We apply the general theory of approximation orders of shift-invariant spaces of [C. de Boor, R. A. DeVore, and A. Ron, *Trans Amer. Math. Soc.*, 341 (1994), pp. 787–806], [C. de Boor, R. A. DeVore, and A. Ron, *J. Funct. Anal.*, 119 (1994), pp. 37–78], and [C. de Boor, R. A. DeVore, and A. Ron, "Approximation orders of FSI spaces," *Constr. Approx.*, 13 (1997), to appear] to the special case when the finitely many generators $\Phi \subset L_2(\mathbb{R}^d)$ of the underlying space S satisfy an N-scale relation (i.e., they form a "father-wavelet" set). We show that the approximation orders provided by such finitely generated shift-invariant spaces are bounded from below by the smoothness class of each $\psi \in S$ (in particular, each $\phi \in \Phi$), as well as by the decay rate of its Fourier transform. In fact, similar results are valid for refinable shift-invariant spaces that are *not* finitely generated.

Specifically, it is shown that under some mild technical conditions on the scaling functions Φ, approximation order k is provided if either some $\psi \in S$ lies in the Sobolev space W_2^{k-1} or its Fourier transform $\widehat{\psi}(w)$ decays near ∞ like $o(|w|^{1-k})$. No technical side conditions are required if the spatial dimension is $d = 1$, and the functions in Φ are compactly supported.

For the special case of a singleton Φ, our first class of results (which are concerned with the condition $\phi \in W_2^{k-1}$) improve previously known results of Meyer and Cavaretta, Dahmen, and Micchelli.

Key words. wavelets, scaling functions, father wavelet, approximation orders, principal shift-invariant spaces, box splines

AMS subject classifications. Primary, 42C15, 41A25; Secondary 41A15

PII. S0036141095282486

1. Introduction and statement of main results.

In this article, we consider the problem of determining *the approximation orders of refinable shift-invariant subspaces* of $L_2 := L_2(\mathbb{R}^d)$, $d \geq 1$. By definition, a subspace $S \subset L_2$ is *shift invariant* (SI) if it is invariant under all *shifts*, i.e., integer translations. We discuss only SI spaces that are *closed* (in L_2). The shift-invariant space S is usually defined with the aid of a *generating set* $\Phi \subset L_2$: we say that Φ *generates* S (and write $S = S(\Phi)$) if S is the smallest (closed) SI space that contains Φ. A *finitely generated SI* (FSI) space is a space $S(\Phi)$ generated by a *finite* Φ, and a *principal SI* (PSI) space $S(\phi)$ is a space generated by a singleton ϕ. PSI and FSI spaces play a role in the theory and applications of multivariate splines, radial-basis function approximation, sampling theory, wavelets, and uniform subdivision schemes. The setup and problem addressed in this paper is particularly relevant to the two latter areas.

In all the above-mentioned applications, the SI space serves as a potential source of approximants. In this regard, then, it becomes important to analyze its "approximation power," preferably in terms of properties of its given generators. One convenient quantitative measurement of this "approximation power" (and the most standard one) is via the notion of *approximation orders*, defined with respect to a

*Received by the editors March 3, 1995; accepted for publication (in revised form) January 29, 1996. This work was partially supported by National Science Foundation grant DMS-9224748, Army Research Office contract DAAH04-95-1-0089, and Israel–U.S. Binational Science Foundation grant 9000220.

http://www.siam.org/journals/sima/28-3/28248.html

[†]Computer Science Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706 (amos@cs.wisc.edu).

ladder of spaces. In the present context, the simplest ladder associated with S is the following *stationary* one.

DEFINITION 1.1. *The* stationary SI ladder generated by Φ *is the directed family*

$$\mathcal{S} := \mathcal{S}(\Phi) := (S_h := \sigma_h S(\Phi))_{h>0},$$

where $S := S(\Phi)$ is the SI space generated by Φ and where σ_h the dilation operator

$$\sigma_h : f \mapsto f(\cdot/h).$$

Note that S_h is "spanned" by the $h\mathbb{Z}^d$-shifts of the dilated functions $\sigma_h\Phi$. (The adjective "stationary" here refers to the fact that finer spaces S_h are obtained from S_1 by dilation. Nonstationary ladders are obtained if one allows each S_h to be spanned by the $h\mathbb{Z}^d$-shifts of some functions $\Phi_h \neq \sigma_h\Phi$.)

DEFINITION 1.2. *Let $\mathcal{S} = (S_h)_h$ be an SI ladder generated by $\Phi \subset L_2$. We say that $\mathcal{S}$* provides approximation order k *to the function space $F \subset W_2^k$ if for every $f \in F$,*

$$\operatorname{dist}(f, S_h) = O(h^k),$$

where "dist" *is the usual L_2-distance between a function and a function set. If the stronger assumption*

$$\operatorname{dist}(f, S_h) = o(h^k)$$

holds, we say that $\mathcal{S}$ provides density order k.

Here and henceforth, W_2^k with k positive is the usual Sobolev/potential space; i.e., if k is an integer, this is the space of all L_2-functions whose weak derivatives up to order k inclusive are in L_2.

The literature concerning L_2- (and, more generally, L_p-) approximation orders of PSI, FSI, and general SI spaces is vast, and reviewing that literature to any extent is not within the scope of this paper. We refer the reader to the introductions and bibliography of [BR], [BDR1], [BDR2], [BDR3], [Jo1], and [Jo2].

The SI ladders that are employed in the context of (the multiresolution approximation approach to) wavelets satisfy an additional important property of refinability.

DEFINITION 1.3. *An SI ladder $\mathcal{S} = (S_h)_h$ is* refinable (*or, more explicitly, N-refinable*) *if for some $N > 1$ and for every integer j,*

$$S_{N^{-j}} \subset S_{N^{-j-1}}.$$

Note that if the ladder is stationary, refinability is implied by the single relation $S_1 \subset S_{1/N}$, and it implies the relation $S_h \subset S_{h/N}$ for all h.

While, in general, the approximation orders of, say, stationary SI ladders $\mathcal{S}(\Phi)$ are unrelated to the smoothness of the generator(s) Φ, it is a known phenomenon that such a relation exists for certain stationary *refinable* PSI ladders. (It is worth mentioning that in the case of a univariate refinable PSI ladder, approximation orders can sometimes be equivalently described in terms of the vanishing moment conditions of the corresponding wavelet.) Results along these lines were proved by people interested in constructing wavelets via multiresolution, as well as by people studying uniform subdivision schemes.

As an example for the former, the following result can be obtained by combining [M, Theorem 2.4] with the quasi-interpolation argument. (Warning: The more explicit result, [M, Theorem 2.6], cannot imply the full approximation order asserted below.)

RESULT 1.4 (see Meyer [M]). *Let k be a positive integer and $\phi \in W_2^{k-1}$. Assume that all derivatives of ϕ of order $< k$ are bounded and rapidly decaying. If the shifts of ϕ are L_2-stable and if the stationary PSI ladder $\mathcal{S}$ generated by ϕ is 2-refinable, then $\mathcal{S}$ provides approximation order k for W_2^k.*

In the context of subdivision, the following result was established by Cavaretta, Dahmen, and Micchelli.

RESULT 1.5 (see [CDM]). *Assume that ϕ is a compactly supported function in $C^{k-1}(\mathbb{R}^d)$, that $\mathcal{S}(\phi)$ is 2-refinable, that the refinement mask of ϕ is finite, and that the underlying subdivision scheme converges uniformly. Then $\mathcal{S}(\phi)$ provides approximation order k for all sufficiently smooth functions.*

We note that the above C^{k-1}-assumption on ϕ appears to be too restrictive. For example, refinable polynomial B-splines as well as refinable polynomial box splines that are not in C^{k-1} (but are in C^{k-2}) provide approximation order k (cf. [BHR] for details). In any event, both results are restricted to PSI spaces and impose fast decay rates on the generator ϕ, together with some kind of stability assumption on its shifts. (We have not defined the notions of L_2-stability or the uniform convergence of subdivision schemes. We do note that the former implies the latter and both imply that $\widehat{\phi}(0) \neq 0$.) We add here that [JM, Theorem 2.4] can be combined with [BDR1, Theorem 1.15] to show that for the specific value $k = 1$, Result 1.4 is valid under very mild decay conditions on ϕ.

Note that all of the above-quoted results show that smoothness may imply approximation orders, and none of them addresses the converse statement. Indeed, there are various examples of high approximation orders of stationary refinable PSI ladders generated by functions of low smoothness (e.g., Daubechies' scaling functions; cf. [D]).

In this paper, we shall establish results concerning the connection between the existence of smooth functions in the refinable S and the corresponding approximation order provided by the ladder $\mathcal{S}$. These results improve their literature counterparts in several ways: First and foremost, they apply to FSI and even to arbitrary SI spaces, while the above-stated results are confined to PSI spaces only. Second, they hardly require any stability or related assumptions on the shifts of a generating set Φ, and they make no assumptions about a possible direct relation between Φ and their dilates. (Recall that the [CDM] result, for example, assumes the relevant mask to be finite.) Third, the results apply to functions Φ that decay only mildly; in fact, the most general results here do not even mention generating sets. Fourth, the results do not require the *generator(s)* to be smooth but only that S contains one smooth function. That latter difference is critically important for SI spaces which are not principal. Further, none of our results restricts the integer value of N in the refinement condition; some of the results do not even require N to be an integer. Finally, the results of this paper remain valid if, instead of assuming that S is refinable and contains smooth functions, we drop the refinability and assume only that $\cap_{i=1}^{\infty} S_{1/N^i}$ contains such smooth function.

The techniques employed also apply to L_p-approximation orders, $p \neq 2$, as well as to nonstationary ladders. However, these extensions will be discussed elsewhere.

We have just mentioned that our results require S to contain smooth functions. We actually use three different conditions to describe "smoothness," only one of which is truly a smoothness condition. We refer to the other two as *pseudosmoothness* *conditions.* The first assumption is that for some k, $W_2^k \cap S \neq 0$. The other two criteria are in terms of the decay of the Fourier transform $\widehat{f}$ of f. Precisely, for some small neighborhood B of the origin, our pseudosmoothness conditions will require the existence of $f \in S$ whose corresponding sequence

$$\lambda_f(m) := \left\| 1 - \frac{|\widehat{f}|^2}{\sum_{j\in 2\pi m\mathbb{Z}^d} |\widehat{f}(\cdot + j)|^2} \right\|_{L_1(B)}^{1/2}, \tag{1.1}$$

or

$$\widetilde{\lambda}_f(m) := \left\| 1 - \frac{|\widehat{f}|^2}{\sum_{j\in 2\pi m\mathbb{Z}^d} |\widehat{f}(\cdot + j)|^2} \right\|_{L_\infty(B)}^{1/2}, \tag{1.2}$$

decays at a certain rate. In the above expressions, $0/0 := 0$.

Discussion. Under mild conditions (e.g., $|\widehat{f}| \geq c >$ a.e. around the origin), there exists a function $g \in S(f)$ such that the decay rate of λ_f at ∞ is the same as the decay rate of the sequence

$$\mu_g : m \mapsto \left\| \sum_{j\in 2\pi m\mathbb{Z}^d\backslash 0} |\widehat{g}(\cdot + j)|^2 \right\|_{L_1(B)}^{1/2}.$$

In particular, if $g \in W_2^k$, then $\mu_g(m) = o(m^{-k})$, and hence also $\lambda_f(m) = o(m^{-k})$. Similarly, if $\widehat{g}(w) = O(|w|^{-k})$, $k > d/2$ (as $|w| \to \infty$), then both $\lambda_f(m)$ and $\widetilde{\lambda}_f(m)$ are $O(m^{-k})$. The discussion thus explains the point of choosing the terminology "pseudosmoothness."

The most general result proved in this paper is as follows.

THEOREM 1.6. *Let $\mathcal{S} = (S_h)_h$ be a stationary SI ladder, and let k' be a positive number. Assume that the following conditions hold:*

(a) *"Pseudorefinability": for some integer N, $S_0 := \cap_{i=1}^{\infty} S_{1/N^i} \neq 0$.*

(b) *"Pseudosmoothness": there exists $\psi \in S_0$ such that with $\widetilde{\lambda}_\psi$ defined as in (1.2), $\widetilde{\lambda}_\psi(N^j) = O(N^{-jk'})$.*

Then $\mathcal{S}$ provides density order k for W_2^k for every $k < k'$.

The highlight of this result lies in the fact that its requirements are almost merely the two basic ones: *pseudorefinability* and *pseudosmoothness.* The space S can be a PSI, FSI, or arbitrary SI space, and the ability to find a "good" generating set for this space is not an issue. The pseudosmoothness assumption on ψ in Theorem 1.6 is satisfied if $\widehat{\psi}$ decays at ∞ like $O(|\cdot|^{-k'})$, provided that $k' > d/2$ and that $1/\widehat{\psi}$ is essentially bounded around the origin. This latter "side condition" (i.e., the essential boundedness of $1/\widehat{\psi}$) cannot be dispensed with: approximation orders of the refinable $\mathcal{S}$ to W_2^k are *not* implied by the mere existence of smooth functions ψ in the refinable S. Here is a simple example.

Example: A refinable analytic PSI space that provides zero approximation order. Let S be the space of all univariate band-limited functions with band in $[0\,..\,2\pi]$. S is a PSI space, and all functions in S are analytic and hence smooth. Moreover, S contains an abundance of rapidly decaying functions. However, the existence of smooth rapidly decaying functions in S cannot be converted to positive assertions concerning approximation orders provided by S: while $\mathcal{S}$ provides very good approximation orders to *some* smooth functions, it provides zero approximation order to many others.

Consequently, already in the PSI context, refinability and smoothness alone cannot ensure good approximation properties. Note that, indeed, all decaying functions (say, L_1) in S must have a zero mean value.

It is quite safe to conjecture that the SI ladder $\mathcal{S}$ in the theorem provides approximation order k' and not only density orders $k < k'$.

The weak aspect of Theorem 1.6 is that the approximation order concluded in this result is bounded above by the "pseudosmoothness parameter" k' of ψ. In comparison, in all other results of this paper, the asserted approximation order will be obtained by "rounding up" this parameter to the next integer. However, it is impossible to achieve such results without further assumptions for the simple reason that there exist refinable spaces whose corresponding (maximal) approximation order is fractional.

We are now ready to present additional selected results from this paper. When reading these subsequent results, it will be convenient to classify the conditions assumed in them as follows: (a) (pseudo)refinability, (b) (pseudo)smoothness, and (c) extra "side conditions."

The first result is a special case of Theorem 3.2.

COROLLARY 1.7. *Let $\mathcal{S}(\Phi)$ be a stationary FSI ladder, and let k be a positive integer. Assume the following:*

(a) *$S(\Phi)$ is refinable: $S_1 \subset S_{1/N}$.*

(b) *$S(\Phi) \cap W_2^{k-1} \neq 0$.*

(c) *The (finite) generating set Φ satisfies the following three "side conditions":*

(c1) *$|\phi(x)| = O(|x|^{-\rho})$ (as $x \to \infty$) for some $\rho > k + d$ and for every $\phi \in \Phi$.*

(c2) *$\widehat{\phi}(0) \neq 0$, for some $\phi \in \Phi$.*

(c3) *The functions $(\sum_{\alpha \in \mathbb{Z}^d} \phi(\cdot - \alpha))_{\phi \in \Phi}$ are linearly independent.*

Then $\mathcal{S}(\Phi)$ provides approximation order k for W_2^k.

The complementary result, which invokes a pseudosmoothness assumption, is the following corollary of Theorem 3.4.

COROLLARY 1.8. *Let $\mathcal{S}(\Phi)$ be a stationary FSI ladder, and let k be a positive integer. Assume the following:*

(a), (c) *These conditions are as in Corollary 1.7, but here N is assumed to be an integer.*

(b) *For some $\psi \in S(\Phi)$, $\lambda_\psi(N^j) = o(N^{-j(k-1)})$, where λ_ψ is defined by (1.1).*

Then $\mathcal{S}(\Phi)$ provides approximation order k for W_2^k.

We remark that for a PSI space $S(\phi)$, condition (c3) is redundant since it is implied by (c2): for a PSI space $S(\phi)$, (c3) is violated if and only if $\sum_{\alpha \in \mathbb{Z}^d} \phi(\cdot - \alpha) = 0$, and it is well known that that can happen only if $\widehat{\phi}(0) = 0$. Thus for the PSI case, Corollaries 1.7 and 1.8 lead to the following result.

COROLLARY 1.9. *Let $S(\phi)$ be a PSI stationary ladder, and let k be a positive integer. Assume the following:*

(a) *$S(\phi)$ is refinable: $S_1 \subset S_{1/N}$.*

(b) *Either of the following two conditions holds:*

(b1) *$S(\phi) \cap W_2^{k-1} \neq 0$.*

(b2) *N is an integer, and $\lambda_\phi(N^j) = o(N^{-j(k-1)})$.*

(c) *The generating function ϕ satisfies the following two "side conditions":*

(c1) *$|\phi(x)| = O(|x|^{-\rho})$ (as $x \to \infty$) for some $\rho > k + d$.*

(c2) *$\widehat{\phi}(0) \neq 0$.*

Then $\mathcal{S}(\phi)$ provides approximation order k for W_2^k.

Discussion. Corollary 1.9 implies that a refinable PSI space $\mathcal{S}(\phi)$ provides approximation order k in the case where $\phi \in W_2^{k-1}$, and the side conditions (c1) and (c2) are met. A comparison of these side conditions with those assumed in Results 1.4 and 1.5 shows that this corollary requires less decay of ϕ and frees the shifts of ϕ almost completely from any "stability" requirements (other than the basic assumption $\widehat{\phi}(0) \neq 0$). In terms of its smoothness requirement, it assumes about the same as Result 1.4 and less than Result 1.5.

Remark. At a late stage, when susbstantial modifications of this paper became prohibitive, R.-Q. Jia brought to my attention the fact that, while I quote [CDM, Theorem 8.3], there is a significantly stronger result there, viz, Theorem 8.4. Indeed, that theorem seems to be on par with the (b1) variant of Corollary 1.9. One should note that our general PSI result (Theorem 3.4 + Proposition 4.1) applies to generating functions of slow decay (such as the sinc function).

Example: B-splines and box splines. The simplest example of a refinable ϕ is the univariate B-spline of order k. It is compactly supported and nonnegative and hence trivially satisfies condition (c) of Corollary 1.9. It is well known that $\mathcal{S}(\phi)$ provides approximation order k. That result is indeed reproduced (twice) by Corollary 1.9. First, $\widehat{\phi}(w) = w^{-k}\tau(w)$ for a certain trigonometric polynomial τ, and hence $\widehat{\phi} = o(|\cdot|^{-k+\varepsilon})$ for any $\varepsilon > 0$, and hence the sequence λ_ϕ also decays at that rate. Second, ϕ can be shown to lie in W_2^s for any $s < k - 1/2$. The situation for box splines (cf. [BHR] for definition and details) is similar.

The B-spline example also shows the sharpness of the pseudosmoothness condition (b2) in Corollary 1.9: the B-spline ϕ of order $k-1$, whose ladder does *not* provide approximation order k, satisfies the condition $\lambda_\phi(m) = O(m^{-(k-1)})$. Corollary 1.9 thus fails to hold if we change the small o in (b2) to big O.

Example: Band-limited functions. Let S be the PSI space of all univariate L_2-functions whose Fourier transform is supported on $[-\pi\,..\,\pi]$. That space is well known to provide all positive approximation orders. Result 1.4 cannot reproduce this fact since S can be easily shown to contain no L_1-function ϕ whose shifts are L_2-stable. In contrast, the space contains an abundance of functions that decay rapidly together with all their derivatives and have nonzero mean value. Hence either one of (b1) and (b2) of Corollary 1.9 can be activated to yield these known spectral orders of approximation. Moreover, we remark that our more general result, Theorem 3.4, requires only the existence of $\psi \in S$ whose Fourier transform is C^∞ around the origin and does not vanish there.

Our next result deals with *univariate FSI ladders generated by compactly supported functions.* This case, though very special, is of much practical interest. The point of the theorem is that in the univariate case, "compact support" is already a "sufficient side condition."

THEOREM 1.10. *Let $\mathcal{S}(\Phi) = (S_h)_h$ be a stationary FSI ladder, and let k be a positive integer. Assume the following:*

(a) *For some $N > 1$, $S_0 := \cap_{i=1}^{\infty} S_{1/N^i} \neq 0$.*

(b) *Either of the following two conditions holds:*

(b1) *$S_0 \cap W_2^{k-1} \neq 0$.*

(b2) *N is an integer and there exists $\psi \in S_0$ such that the sequence λ_ψ defined in* (1.1) *satisfies $|\lambda_\psi(N^j)| = o(N^{-j(k-1)})$.*

(c) *The spatial dimension d is* 1, *and Φ are compactly supported.*

Then $\mathcal{S}(\Phi)$ provides approximation order k for W_2^k.

Example: C^1-cubics. This example is taken from [HSS]. Let S be the space of all univariate piecewise-cubic polynomials with breakpoints at the integers and which are globally C^1. This space is obviously N-refinable (for all integers N). Less obviously, but as is quite well known, it is a local FSI space of length 2, i.e., it is generated by two compactly supported functions. It is fairly obvious that the approximation order here is 4. (The subspace of C^2-cubics already does the job.) That order is recovered from Theorem 1.10 as soon as we realize that the (smooth) B-spline of order 4 is in our space. In contrast, standard generating sets for this space consist of functions each of which neither lies in W_2^3 nor satisfies the alternative requirement, (b2), that appears in Theorem 1.10. Thus it is very important that our results are stated in terms of the smoothness of some function *in the space* and *not* in terms of the smoothness of some function *in the generating set.*

Discussion. The example given after Theorem 1.6 shows that "compact support" in Theorem 1.10 cannot be replaced by "rapid algebraic decay." However, the theorem *is* extendible to exponentially decaying generators.

All results as stated aim at providing *lower* bounds on the approximation order of the ladder $\mathcal{S}(\Phi)$ in terms of either the smoothness of the "smoothest" function in $S(\Phi)$ or the decay of its Fourier transform. Such presentation stems from the typical problem in *spline theory,* where smoothness is a more readily available property than the approximation orders (cf. [BHR]). However, for more general refinable functions, the readily available information (viz, the mask) may appear to be more adequate for computing approximation orders than estimating either the smoothness of, say, $\phi \in \Phi$ or the decay of its Fourier transform. From this point of view, the results in this paper can be regarded as providing *upper* bounds on the possible smoothness of functions in $S(\Phi)$ in terms of the known approximation order of the underlying ladder. As an illustration, we state the following immediate corollary.

COROLLARY 1.11. *Let $S(\Phi)$ be a univariate refinable FSI space generated by compactly supported functions. Then no compactly supported $\psi \in S(\Phi)$ is infinitely many times continuously differentiable.*

Proof. Let $\psi \in S(\Phi)$ be nonzero, compactly supported, and C^∞. Then $\psi \in W_2^k$ for *every* k. By Theorem 1.10, $\mathcal{S}$ then provides all positive approximation orders. By [J, Theorem 5], the shifts of Φ must then span all polynomials, an absurdity in view of the fact that these shifts have finite local dimension. □

The argument extends to more than one variable. One only needs to adopt further conditions. If Φ is a singleton (ϕ), the additional condition is that $\widehat{\phi}(0) \neq 0$ (so that we will be able to apply Corollary 1.9). For a finite Φ, the conditions should be (c2) and (c3) of Corollary 1.7.

As alluded to before, there is a tight connection between the approximation orders of the refinable $\mathcal{S}$ and the vanishing moments of the wavelets. Many of our results can thus be stated in terms of such vanishing moments. Here is one illustration.

COROLLARY 1.12. *Let $\mathcal{S} = (S_h)_h$ be a univariate stationary FSI ladder generated by compactly supported functions, and let k be a positive integer. Assume further that for some $N > 1$, $\cap_{i=1}^{\infty} S_{1/N^i} \cap W_2^{k-1} \neq 0$. Let f be compactly supported. If $f \perp S$, then $\widehat{f}$ has a zero of order k at the origin.*

Proof. From Theorem 1.10, we conclude that $\mathcal{S}$ provides approximation order k. [BDR2, Theorem 4.1] then applies to yield that for some compactly supported supported $\phi \in S$, the ladder $\mathcal{S}(\phi)$ also provides approximation order k. [R2, Theorem 3.7], when combined with [R1, Theorem 1.1], allows us to assume without loss of generality that $\widehat{\phi}$ has no 2π-periodic zero. On the other hand, [BDR1, Theorem 1.14] implies that $\widehat{\phi}$ must have a zero of order k at each $j \in 2\pi\mathbb{Z}\backslash 0$. This forces $\widehat{\phi}(0)$ to be nonzero, and by a standard argument, we may assume that $\widehat{\phi} - 1$ has a k-fold zero at the origin.

The rest of the argument is routine. Let f be a compactly supported L_2-function, and set

$$H := \sum_{j \in 2\pi\mathbb{Z}} \widehat{f}(\cdot + j)\overline{\widehat{\phi}(\cdot + j)}.$$

In general, the above sum is L_1-convergent. However, since f and ϕ are compactly supported, one can show that the above sum can be differentiated term by term and that each such differentiated sum converges uniformly on compactly sets. In view of the properties of $\widehat{\phi}$, this implies that $H - \widehat{f}$ has a zero of order k at the origin. Assuming in addition that $f \perp S(\phi)$ (certainly true if $f \perp S$), Poisson's summation formula yields that $H = 0$ and hence that $\widehat{f}$ indeed has a zero of order k at the origin. □

Finally we comment on the proofs of the main results. Theorem 3.2 and its corollary follow as a strikingly simple consequence of the general theory of [BDR1]. A totally different (and somewhat tricky) argument is employed in the proofs of Theorems 1.6 and 3.4; the latter argument leads to further consequences, which will be discussed elsewhere. The two approaches differ also from a conceptual point of view. In the first approach, the function $f \in S \cap W_2^{k-1}$ is *approximated by* the ladder $\mathcal{S}$. In the second approach, the pseudosmooth function ψ whose Fourier transform decays nicely is used *to provide approximants* to other smooth functions.

The paper is organized as follows. In section 2, we collect general results concerning approximation orders of stationary SI ladders. None of the results of section 2 assume refinability, and all results should be considered "auxiliary" from the standpoint of this paper. The three basic results of this paper—the proof of Theorem 1.6, Theorem 3.2, and Theorem 3.4—comprise section 3. The latter theorems of section 3 assume a new, initially obscure property of the space S, the $H(k)$ property. The study of this property is presented in section 4, where we show that PSI and FSI spaces with "reasonably good" generating sets satisfy it. (The results in section 4 do not rely on refinability, and refinability buys no extra benefit; hence we decided to separate this discussion from that of section 3.)

2. Background on the approximation order of stationary SI ladders.

In this section, we collect all known and new results on approximation from stationary SI ladders that are used as *auxiliary* results in this paper. We emphasize that none of these results use the *refinability* assumption, which is the pillar assumption in the main results of this article.

The following function plays the key role in the determination of the approximation order of the stationary PSI ladder $\mathcal{S}(\phi)$:

$$\Lambda_\phi := \left(1 - \frac{|\widehat{\phi}|^2}{\sum_{j \in 2\pi\mathbb{Z}^d} |\widehat{\phi}(\cdot + j)|^2}\right)^{1/2}. \tag{2.1}$$

In this definition, $0/0 := 0$.

The basic observation of [BDR1] is the following.

(2.2) [BDR1, THEOREM 2.20]. *Let f be a function whose Fourier transform is supported in the cube C/h, $C := [-\pi \,..\, \pi]^d$. Let $\phi \in L_2$. Then*

$$\operatorname{dist}(f, \sigma_h S(\phi)) = (2\pi)^{-d/2} \|\Lambda_\phi(h\cdot)\widehat{f}\|_{L_2(C/h)}.$$

Whenever $\widehat{f}$ is not supported on C/h, the function f is split in [BDR1] into f_1+f_2, where $\widehat{f_1}$ coinciding with $\widehat{f}$ on B/h, $B \subset C$, and is zero elsewhere. This leads (almost immediately) to the following estimates.

COROLLARY 2.1. *Let $\mathcal{S}(\phi) = (S_h)_h$ be a stationary PSI ladder, and let B be a small neighborhood of the origin. Then for every $f \in L_2$ and every $h > 0$,*

$$\operatorname{dist}(f, S_h) \leq (2\pi)^{-d/2}(\|\Lambda_\phi(h\cdot)\widehat{f}\|_{L_2(B/h)} + \|\widehat{f}\|_{L_2(\mathbb{R}^d\backslash(B/h))}).$$

In particular, if $f \in W_2^k$, then

$$\operatorname{dist}(f, S_h) = (2\pi)^{-d/2}\|\Lambda_\phi(h\cdot)\widehat{f}\|_{L_2(B/h)} + o(1)h^k\|f\|_{W_2^k},$$

where the $o(1)$ expression is bounded independently of f.

From that, the following characterization is provided in [BDR1].

(2.3) [BDR1, THEOREM 1.6]. *The following conditions are equivalent for any $k > 0$:*

(a) *The function $|\cdot|^{-k}\Lambda_\phi$ is in $L_\infty(B)$ for some 0-neighborhood B.*

(b) *The stationary PSI ladder $\mathcal{S}(\phi) = (S_h)_h$ provides approximation order k in the strong sense that*

$$\operatorname{dist}(f, S_h) \leq \operatorname{const}\|f\|_{W_2^k}h^k \tag{2.4}$$

for some const *independent of f and h.*

Under some favorable conditions on the generator ϕ, one is able to derive simpler characterizations from the last one. One such simplification in contained in the following statement.

(2.5) [BDR1, COROLLARY 5.15]. *Assume that ϕ satisfies assumption* (c) *of Corollary* 1.9. *Then $\mathcal{S}(\phi)$ provides approximation order k if and only if ϕ satisfies the Strang–Fix conditions of order k, i.e., $\widehat{\phi}$ has a zero of order k at each $j \in 2\pi\mathbb{Z}^d\backslash 0$.*

The treatment of the approximation orders of FSI and SI spaces here is exclusively based on the powerful "superfunction" theory that surrounds the existence of a "superfunction" $\psi \in S$, i.e., a function ψ whose corresponding $\mathcal{S}(\psi)$ provides the same approximation order as the larger ladder $\mathcal{S}$. An overview of the superfunction results in the literature prior to 1991 can be found in [BDR1, section 1]. [BDR1, section 3], [BDR2, section 4], and all of [BDR3] constitute the most advanced progress on this problem.

(2.6) THE STATIONARY CASE OF THE "SUPERFUNCTION" RESULTS OF [BDR1], [BDR2], AND [BDR3]. *Let $\mathcal{S} = (S_h)_h$ be a stationary SI ladder, and let $k > 0$. Then for any $\chi \in L_2$, there exists a stationary PSI ladder $\mathcal{T} = (T_h)_h$ such that $T_1 \subset S_1$ and*

(a) *For every $f \in L_2$,*

$$\operatorname{dist}(f, T_h) \leq \operatorname{dist}(f, S_h) + 2\operatorname{dist}(f, \sigma_h S(\chi)).$$

(b) *If S is finitely generated by compactly supported functions and χ is compactly supported, T is generated by some compactly supported function.*

(c) *If $\mathcal{S}(\chi)$ provides approximation order $k+1$ to W_2^{k+1}, then for every $f \in W_2^k$,*

$$\operatorname{dist}(f, T_h) \leq \operatorname{dist}(f, S_h) + \varepsilon_f(h)h^k\|f\|_{W_2^k},$$

where ε_f is bounded independently of f and h and decays to 0 with h.

(d) *If S provides approximation order k in the sense that*

$$\operatorname{dist}(f, S_h) \leq \text{const}\|f\|_{W_2^k} h^k \quad \forall f \in W_2^k,$$

then it provides density order k' for every $f \in W_2^{k'}$ and every $k' < k$.

Proof. (a) follows from [BDR1, Theorem 3.3]. (b) is the content of [BDR2, Theorem 4.1]. We prove (c) and (d) simultaneously as follows. First, the PSI case of (d) follows from a simple comparison of the characterization of approximation orders [BDR1, Theorem 1.6] (see above) and the characterization of density orders [BDR1, Theorem 1.7]. (c) then follows by an application of (d) to the PSI ladder $\mathcal{S}(\chi)$, with $k+1$ replacing k.

It then remains to show that (d) is valid for a general SI space: assuming that $\mathcal{S}$ satisfies the assumption in (d), we choose χ such that $\mathcal{S}(\chi)$ provides approximation order $k+1$, and we then let $\mathcal{T}$ be the PSI ladder of (a) with respect to the current χ. By (c), the ladder $\mathcal{T}$ provides approximation order k (in the sense required in (d)). Since $\mathcal{T}$ is principal and (d) is known to be valid with respect to principal ladders, we conclude that $\mathcal{T}$ provides density orders $k' < k$; this is true a fortiori for the original ladder $\mathcal{S}$. □

Finally, the following result, which is needed for the proof of Theorem 1.6, cannot be found in the literature.

THEOREM 2.2. *Let $\mathcal{S}$ be a stationary SI ladder, let $m > k > 0$, and let $(h_i)_i$ be decreasing to zero and satisfy*

$$h_i/h_{i+1} \leq A < \infty \quad \forall i.$$

Assume that for every $f \in W_2^m$,

$$\operatorname{dist}(f, S_{h_i}) \leq c h_i^k \|f\|_{W_2^m}, \tag{2.7}$$

with c independent of f (and h). Then $\mathcal{S}$ provides approximation order k for W_2^k in the sense of (2.4).

Proof. Invoking (2.7) together with (2.6)(c) (with k there being our m here), we find $\phi \in S$ such that

$$\operatorname{dist}(f, \sigma_{h_i} S(\phi)) \leq c h_i^k \|f\|_{W_2^m}.$$

This reduces the problem to the PSI ladder $\mathcal{S}(\phi)$ since $\mathcal{S}$ certainly provides approximation order k the moment that $\mathcal{S}(\phi)$ does. Therefore, without loss of generality, we may assume that $S = S(\phi)$, i.e., that our ladder is principal.

We now invoke Corollary 2.1 (with k there being our m here) to conclude that (2.7) implies that

$$\|\Lambda_\phi(h_i\cdot)\widehat{f}\|_{L_2(B/h_i)} \leq c h_i^k \|f\|_{W_2^m}.$$

Since f varies over W_2^m, $(1+|\cdot|)^{2m}\widehat{f}^2$ varies over L_1, and we may then convert the last inequality to

$$\|(1+|\cdot|)^{-2m}\Lambda_\phi(h_i\cdot)^2 f\|_{L_1(B/h_i)} \leq c h_i^{2k}\|f\|_{L_1} \quad \forall f \in L_1.$$

This implies the estimate

$$h_i^{2m}\|(h_i+|\cdot|)^{-2m}\Lambda_\phi^2\|_{L_\infty(B)} = \|(1+|\cdot|)^{-2m}\Lambda_\phi(h_i\cdot)^2\|_{L_\infty(B/h_i)} \leq c h_i^{2k}.$$

Assuming that $h_{i+1} \leq |x| \leq h_i$, we obtain

$$\Lambda_\phi(x)^2 \leq c' h_i^{2k} \leq c' A^{2k} |x|^{2k}.$$

Thus the function $| \cdot |^{-k} \Lambda_\phi$ was proved to be bounded around the origin, which in view of [BDR1, Theorem 1.6] (cf. (2.3)) implies that $\mathcal{S}$ provides approximation order k to all of the functions in W_2^k. □

3. Core. Let S be an SI space. The main results of this paper are based on the presumption that the ladder $\mathcal{S}$ can provide approximation order k to either "almost no" smooth functions or "almost all" smooth functions. The most extreme statement of this nature is contained in the following definition.

DEFINITION 3.1. *Let $\mathcal{S} = (S_h)_h$ be a stationary (not necessarily refinable) SI ladder. Let k be a positive integer. We say that $\mathcal{S}$ has* property H(k) *if it provides approximation k for the entire W_2^k, whenever the following condition holds: There exists a function $f \in W_2^{k-1} \backslash 0$ such that for some sequence $(h_i)_i$ that decreases to zero,*

$$\operatorname{dist}(f, S_h) = o(h^{k-1}), \quad h = h_1, h_2, \ldots.$$

In the language of [BDR1], the satisfaction of property H(k) implies that $\mathcal{S}$ provides *approximation* order k to all reasonably smooth functions as soon as it provides a *density* order $k-1$ to a single *smooth* nonzero function.

The main branch of our approach can be now clearly stated: we will show that "smoothness implies approximation orders" for SI ladders $\mathcal{S}$ that are *refinable* and *satisfy the* H(k) *property.* The two theorems that establish this fact are stated and proved in this section, and they apply to an arbitrary SI ladder (i.e., not an FSI one).

Of course, such results may be deemed useless unless we are able to find feasible side conditions on S that guarantee the satisfaction of property H(k). This complementary study is independent of the refinability or smoothness assumptions and is even independent of the possible approximation order provided by $\mathcal{S}$. In fact, we show that some "mild technical conditions" (which are known to neither imply nor be implied by nor be related in any rigorous way to approximation orders) guarantee the satisfaction of property H(k). *However, this complementary analysis does require our space S to be a PSI or FSI space.*

Our first observation is, actually, trivial.

THEOREM 3.2. *Let k be a positive integer, and let $\mathcal{S} = (S_h)_h$ be a stationary SI ladder. Assume further that*

(a) *for some $N > 1$, $\cap_{i=1}^{\infty} S_{1/N^i} \cap W_2^{k-1} \neq 0$;*

(b) *$\mathcal{S}$ has property* H(k).

Then the ladder $\mathcal{S}$ provides approximation order k for W_2^k.

Proof. Let f be a nonzero function in $\cap_{i=1}^{\infty} S_{1/N^i} \cap W_2^{k-1}$. Setting

$$V_i := S_{N^{-i}} \quad \forall i \geq 0,$$

we know that $f \in V_i$, $\forall i \geq 0$. This means that $\operatorname{dist}(f, V_i) = 0$ for all $i \geq 0$. Thus we may invoke property H(k) with respect to this f to conclude that $\mathcal{S}$ provides approximation order k to W_2^k, as asserted. □

The other two results in this section, the proof of Theorem 1.6 and Theorem 3.4, are very similar each to the other, not only in their pseudosmoothness assumptions but also in their methods of proof. However, they differ quite significantly in their conclusions, and hence Theorems 1.6 and 3.4 have been labeled and situated separately.

Proof of Theorem 1.6. Let $k'' < k'$, and let $k > k''$ be any number of the form $k = (r/(r+1))k'$, where r is an integer. We will prove that there exists a sequence $(h_i = a^i)_i$ $(a < 1)$ and a constant c such that for every $f \in W_2^{(r+1)k}$,

$$\text{dist}(f, S_h) \le c\|f\|_{W_2^{(r+1)k}} h^k, \quad h = h_1, h_2, \ldots. \tag{3.1}$$

Theorem 2.2 would then yield that $\mathcal{S}$ provides approximation order k to W_2^k, and (2.6)(d) would then complete the proof.

In what follows, we set $\widetilde{\lambda} := \widetilde{\lambda}_\psi$ (cf. (1.2)) and prove (3.1). However, since the method here will be needed (with a slight twist) in the proof of Theorem 3.4, we formalize it as a separate lemma.

LEMMA 3.3. *Let $\mathcal{S}$ be a stationary ladder, and let $\psi \in \cap_{i=1}^\infty S_{1/N^i} \backslash 0$, where N is a positive integer. Define the sequences $\lambda := \lambda_\psi$ and $\widetilde{\lambda} := \widetilde{\lambda}_\psi$ as in* (1.1) *and* (1.2). *Let n be a positive power of N, u be a positive number, and $h := u/n$. Then for every $f \in L_2$,*

$$\text{dist}(f, S_h) \le \text{const}(\|\widehat{f}\|_{L_2}\widetilde{\lambda}(n) + \|\widehat{f}\|_{L_2(\mathbb{R}^d\backslash(B/u))})$$

and

$$\text{dist}(f, S_h) \le \text{const}(u^{-d/2}\|\widehat{f}\|_{L_\infty}\lambda(n) + \|\widehat{f}\|_{L_2(\mathbb{R}^d\backslash(B/u))}),$$

with const *depending on d only. Of course, the second estimate is meaningful only when $\widehat{f} \in L_\infty$.*

Proof. We split $f = f_1 + f_2$, with $\widehat{f_1}$ coinciding with $\widehat{f}$ on B/u and $\widehat{f_2}$ coinciding with $\widehat{f}$ on $\mathbb{R}^d\backslash(B/u)$. Since obviously

$$\text{dist}(f, S_h) \le \text{dist}(f_1, S_h) + \|f_2\|$$

and since the Fourier transform is an isometry on L_2, we see that the proof of the theorem is reduced to proving that

$$\text{dist}(f_1, S_h) \le \text{const}\|\widehat{f}\|_{L_2}\widetilde{\lambda}(n) \tag{3.2}$$

and

$$\text{dist}(f_1, S_h) \le \text{const } u^{-d/2}\|\widehat{f}\|_{L_\infty}\lambda(n).$$

We will prove (3.2). The proof will eventually establish the other bound as well.

Let ϕ be any function in $S_{1/n}$. This latter space is invariant under $\mathbb{Z}^d/n$-shifts, and since n is an integer, it is also invariant under integer shifts. Thus we not only have $\phi \in S_{1/n}$ but also $S(\phi) \subset S_{1/n}$. Applying dilation, we obtain that

$$\sigma_u S(\phi) \subset \sigma_u S_{1/n} = S_h.$$

Thus instead of proving (3.2), we are entitled to prove that for some $\phi \in S_{1/n}$,

$$\text{dist}(f_1, \sigma_u S(\phi)) \le \text{const}\|f\|_{L_2}\widetilde{\lambda}(n). \tag{3.3}$$

However, $\widehat{f_1}$ is supported on B/u, and therefore [BDR1, Theorem 2.20] (cf. (2.2)) provides us with the explicit formula

$$\operatorname{dist}(f_1, \sigma_u S(\phi)) = \operatorname{const}\|\Lambda_\phi(u\cdot)\widehat{f_1}\|_{L_2(B/u)},$$

with Λ_ϕ defined as in (2.1). Comparing this with the desired result (3.3) and taking into account the fact that $\|f_1\|_{L_2} \le \|f\|_{L_2}$, we realize that our claim is established as soon as we can find $\phi \in S_{1/n}$ such that

$$\|\Lambda_\phi\|_{L_\infty(B)} \le \widetilde{\lambda}(n). \tag{3.4}$$

(For the complementary case, we need an estimate $\|\Lambda_\phi\|_{L_2(B)} \le \lambda(n)$.) Since $\psi \in S_{N^{-i}}$, $i \ge 1$, and n is a power of N, $\psi \in S_{1/n}$. [BDR1, Theorem 2.14] then entails that any L_2-function ϕ whose Fourier transform is of the form $\widehat{\phi} = \tau\widehat{\psi}$ is in $\sigma_{1/n}S(\psi) \subset S_{1/n}$, provided that τ is $2\pi n$-periodic. We take τ to be the $2\pi n$-periodization of the support function of B and define ϕ accordingly. Then $\widehat{\phi}$ vanishes on each domain of the form $j+B$, $j \in 2\pi(\mathbb{Z}^d\backslash n\mathbb{Z}^d)$, and $\widehat{\phi} = \widehat{\psi}$ on domains of the form $j+B$, $j \in 2\pi n\mathbb{Z}^d$. Therefore, on B,

$$\sum_{j\in 2\pi\mathbb{Z}^d} |\widehat{\phi}(\cdot + j)|^2 = \sum_{j\in 2\pi n\mathbb{Z}^d} |\widehat{\psi}(\cdot + j)|^2.$$

We then conclude that on B,

$$\Lambda_\phi^2 = 1 - \frac{|\widehat{\psi}|^2}{\sum_{j\in 2\pi n\mathbb{Z}^d} |\widehat{\psi}(\cdot + j)|^2}.$$

This shows that $\widetilde{\lambda}(n) = \|\Lambda_\phi\|_{L_\infty(B)}$, thereby implying the desired result. □

We now return to the proof of Theorem 1.6. Remember that we want to show that (3.1) holds. We apply Lemma 3.3 with respect to $n := N^{ir}$ and $u := N^{-i}$, where i is an integer. Since $\widetilde{\lambda}_\psi(n) \le cn^{-k'}$, then with $h_i := N^{-i(r+1)} = u/n$,

$$\operatorname{dist}(f, S_{h_i}) \le \operatorname{const}(n^{-k'}\|f\|_{L_2} + \|\widehat{f}\|_{L_2(\mathbb{R}^d\backslash(N^iB))}). \tag{3.5}$$

We compute that

$$n^{-k'} = N^{-irk'} = N^{-i(r+1)k} = h_i^k,$$

which takes care of the first term in the right-hand side of (3.5). As for the second term, since $f \in W_2^{(r+1)k}$, it is easy to see that

$$\|\widehat{f}\|_{L_2(\mathbb{R}^d\backslash(N^iB))} = o(1)N^{-ik(r+1)}\|f\|_{W_2^{(r+1)k}} = o(1)h_i^k\|f\|_{W_2^{(r+1)k}},$$

with $o(1)$ bounded independently of f. Thus (3.1) follows from (3.5), and the proof is complete. □

THEOREM 3.4. *Let k and N be positive integers, and let $\mathcal{S}$ be a stationary SI ladder. Assume the following:*

(a) *For some $\psi \in \cap_{i=1}^\infty S_{1/N^i}$, $\lambda_\psi(N^j) = o(N^{-(k-1)j})$.*

(b) *$\mathcal{S}$ has property* H(k).

Then $\mathcal{S}$ provides approximation order k for W_2^k.

Proof. To invoke the assumed property H(k), we take f to be any band-limited Schwartz function. We will show that for some sequence $(h_i)_i$,

$$\operatorname{dist}(f, S_{h_i}) = o(h_i^{k-1}).$$

Property H(k) would then yield the approximation order assertion.

We choose $(h_i)_i$ as follows: since $\lambda_\psi(N^j) = o(N^{(1-k)j})$, we have for each integer i and for all sufficiently large j that

$$\lambda_\psi(N^j) \le \frac{1}{i} N^{i(1-k-d/2)} N^{(1-k)j}. \tag{3.6}$$

For each i, we choose j that satisfies (3.6), define

$$h_i := N^{-(i+j)},$$

and invoke Lemma 3.3, with u there being N^{-i} and n there being N^j. Since f is band-limited, we may take i sufficiently large to ensure that $\widehat{f}$ is supported on $N^i B = B/u$. Thus the lemma together with (3.6) provides the estimate

$$\begin{aligned}\operatorname{dist}(f, S_{h_i}) &\le \operatorname{const}|u|^{-d/2}\|\widehat{f}\|_{L_\infty}\lambda_\psi(N^j) \le \operatorname{const}N^{id/2}\frac{1}{i}N^{i(1-k-d/2)}N^{j(1-k)}\\ &= \operatorname{const}\frac{1}{i}h_i^{k-1}.\end{aligned}$$

The last estimate implies the desired estimate

$$\operatorname{dist}(f, S_{h_i}) = o(h_i^{k-1}). \qquad \square$$

4. Property H(k). While two of the three main results of the previous section require $\mathcal{S}$ to satisfy property H(k), some readers may suspect this property to be as complicated and demanding as the notion of approximation orders. (After all, approximation orders appear in its statement.) However, this is not true: basic decay and regularity conditions on the generating set of S, which are totally unrelated to the approximation orders that space may provide, suffice for guaranteeing the satisfaction of property H(k). The section contains three results along these lines: in the first, we treat the PSI case; in the second, we treat the FSI case; and in the last, we treat univariate FSI spaces generated by compactly supported functions. The proofs of these results are postponed until we show how the three results, when combined with Theorems 3.2 and 3.4, yield the various corollaries stated in the introduction.

PROPOSITION 4.1. *Let k be a positive integer and $\phi \in L_2$. Let $\rho > k + d$, and consider the following conditions:*

(a) *$\phi = O(|\cdot|^{-\rho})$ near ∞, and $\widehat{\phi}(0) \neq 0$.*

(b) *Λ_ϕ^2 (defined in (2.1)) is k times continuously differentiable around the origin.*

(c) *The PSI ladder $\mathcal{S}(\phi)$ has property* H(k).

Then (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c).

Proof of Corollary 1.9. We first assume claims (a), (b1), and (c) of the corollary and apply Theorem 3.2. Comparing the assumptions of the theorem to those of the corollary (i.e., (a), (b1), and (c)), we immediately observe that we only need to show that assumption (c) of the corollary implies property H(k). That latter implication follows directly from Proposition 4.1.

The proof of the other case is identical, only here we invoke Theorem 3.4. □

The only difference between the FSI case and its special PSI case is that it is much harder to verify property H(k) in the FSI context. We forgo generalizing Proposition 4.1 completely and prefer instead to focus on its main implication ((a) $\Longrightarrow$ (c)).

PROPOSITION 4.2. *Let $\Phi \subset L_2$ be a finite set that satisfies condition* (c) *of Corollary* 1.7 *with respect to some positive integer k. Then $S(\Phi)$ satisfies property* H(k).

Corollary 1.7 now follows from Theorem 3.2 and Proposition 4.2, while Corollary 1.8 follows from Theorem 3.4 and Proposition 4.2.

Finally, Theorem 1.10 follows from Theorems 3.2 and 3.4 when combined with the following general observation.

PROPOSITION 4.3. *A univariate FSI ladder that is generated by compactly supported functions satisfies property* H(k) *for all integer values k.*

We now turn to the proofs of Propositions 4.1, 4.2, and 4.3.

Proof of Proposition 4.1.

(a) $\Longrightarrow$ (b): Let ρ' be any number between $k+d/2$ and $\rho-d/2$. Since $\rho' < \rho-d/2$, it easily follows from the Plancheral theorem that $\widehat{\phi} \in W_2^{\rho'}$. Since $\rho' > k+d/2$, the Sobolev embedding theorem then implies that

$$\| |\widehat{\phi}|^2 \|_{C^k(j+B)} \leq \text{const} \|\widehat{\phi}\|^2_{W_2^{\rho'}(j+B)}.$$

Summing over all $j \in 2\pi\mathbb{Z}^d$ and using the subadditivity of the $W_2^{\rho'}$-norm (which is valid with respect to a set of disjoint cubes, as is the case here; cf. [A, p.225]), we obtain that

$$\sum_{j\in 2\pi\mathbb{Z}^d} \| |\widehat{\phi}|^2 \|_{C^k(j+B)} \leq \text{const}' \|\widehat{\phi}\|^2_{W_2^{\rho'}(B+2\pi\mathbb{Z}^d)} < \infty.$$

This readily implies that $\sum_{j\in 2\pi\mathbb{Z}^d} |\widehat{\phi}(\cdot + j)|^2$ is k times continuously differentiable on B. Since this function does not vanish at 0 (since $\widehat{\phi}(0) \neq 0$), we finally conclude that $\Lambda_\phi^2 \in C^k(B)$.

(b) $\Longrightarrow$ (c): The proof requires the following elementary lemma.

LEMMA 4.4. *Let $g \in L_2(\mathbb{R}^d)$, and let B be open and bounded. Then as $h \to 0$,*

$$\| |\cdot| g \|_{L_2(B/h)} = o(h^{-1}).$$

Proof. Without essential loss of generality, we may assume that B is the Euclidean unit ball. We fix h and abbreviate $B_1 := B/\sqrt{h}$ and $B_2 := (B/h)\backslash B_1$. Then obviously $B/h = B_1 \cup B_2$, and hence

$$\begin{aligned} \| |\cdot| g \|^2_{L_2(B/h)} &= \| |\cdot| g \|^2_{L_2(B_1)} + \| |\cdot| g \|^2_{L_2(B_2)} \\ &\leq h^{-1} \|g\|^2_{L_2(\mathbb{R}^d)} + h^{-2} \|g\|^2_{L_2(\mathbb{R}^d \backslash B_1)} = O(h^{-1}) + o(h^{-2}) = o(h^{-2}). \end{aligned}$$ □

We now show how to derive (c) from (b). In order to prove that property H(k) holds, we assume that

$$\operatorname{dist}(f, S_h) = o(h^{k-1}), \tag{4.1}$$

with $f \in W_2^{k-1}\backslash 0$, $h = h_1, h_2, \ldots$, and $(h_i)_i$ decreasing to 0. We will show that this assumption, together with assumption (b), implies that $\mathcal{S}(\phi)$ provides approximation order k for all W_2^k. In what follows, h is always selected from the sequence $(h_i)_i$.

First, substituting $k-1$ for k in the second displayed equation of Corollary 2.1 and using (4.1), we obtain

$$\|\Lambda_\phi(h\cdot)\widehat{f}\|_{L_2(B/h)} = o(h^{k-1}). \tag{4.2}$$

We contend that (4.2) forces all derivatives of Λ_ϕ up to order $k-1$ to vanish at the origin. Since a similar argument is needed in the proof of Proposition 4.3, we prove this fact in a separate lemma.

LEMMA 4.5. *Let* M *be any function that is defined on a neighborhood B of the origin and is k times continuously differentiable there. Let $f \in W_2^{k-1}\backslash 0$. If*

$$\|\mathrm{M}(h\cdot)\widehat{f}\|_{L_2(B/h)} = o(h^{(k-1)}), \tag{4.3}$$

then all derivatives of M *up to order $k-1$ must vanish at the origin.*

Proof. Assume to the contrary that some derivatives of order $l < k$ of M do not vanish at the origin, and let l be the minimal integer with that property. Using the Taylor series expansion of M around 0, we find that on B/h we have an estimate of the form

$$|\mathrm{M}(h\cdot)| \geq h^l|p| - ch^{l+1}|\cdot|^{l+1}.$$

Here c depends on ϕ, B, and l but not on h, and p is a homogeneous polynomial of degree l. Thus invoking that estimate, we obtain from (4.3) that

$$\begin{aligned} o(h^{(k-1)}) &= \|\mathrm{M}(h\cdot)\widehat{f}\|_{L_2(B/h)} \\ &\geq h^l\|p\widehat{f}\|_{L_2(B/h)} - ch^{(l+1)}\||\cdot|^{l+1}\widehat{f}\|_{L_2(B/h)}. \end{aligned} \tag{4.4}$$

Since $f \in W_2^{k-1}$ and $l \leq k-1$, we conclude that $g := |\cdot|^l\widehat{f} \in L_2$. Lemma 4.4 then applies to yield that

$$\||\cdot|^{l+1}\widehat{f}\|_{L_2(B/h)} = \||\cdot|g\|_{L_2(B/h)} = o(h^{-1}).$$

Substituting this estimate into (4.4), we finally obtain

$$h_i^l\|p\widehat{f}\|_{L_2(B/h_i)} = o(h_i^l),$$

which can happen only if $p\widehat{f} = 0$, a contradiction. (Since it is in $L_2\backslash 0$, $\widehat{f}$ cannot be supported on the zero set of p since the latter is a null set.) □

We proceed now with the proof of the theorem. Since, as contended, all derivatives of Λ_ϕ of orders $< k$ vanish at the origin, we see that

$$\Lambda_\phi = O(|\cdot|^k)$$

around the origin. Hence [BDR1, Theorem 1.6] (cf. (2.3)) implies that $\mathcal{S}(\phi)$ provides approximation order k to W_2^k. □

Proof of Proposition 4.2. Since we need to prove that property H(k) holds, we assume that for some $f \in W_2^{k-1}\backslash 0$ and a subsequence $(h_i)_i$,

$$\mathrm{dist}(f, \sigma_h\mathcal{S}(\Phi)) = o(h^{k-1}), \quad h = h_1, h_2, \ldots.$$

We then need to show that $\mathcal{S}(\Phi)$ provides approximation order k.

For this, we will find a function ψ in $\mathcal{S}(\Phi)$ that satisfies the following three conditions:

(i) The function

$$\Lambda_\psi^2 := 1 - \frac{|\widehat{\psi}|^2}{\sum_{\alpha \in 2\pi\mathbb{Z}^d} |\widehat{\psi}(\cdot + \alpha)|^2}$$

is k times continuously differentiable around the origin.

(ii) $\operatorname{dist}(f, T_h) = o(h^{k-1})$, $h \in (h_i)_i$, where T_h is the h-dilate of $S(\psi)$.

In view of Proposition 4.1, condition (i) implies that $\mathcal{S}(\psi)$ satisfies property H(k); hence, in view of condition (ii), it must provide approximation order k (for all functions in W_2^k). Since $S(\psi) \subset S(\Phi)$, $\mathcal{S}(\Phi)$ must then provide this approximation order as well.

We need now to prove that ψ as above exists, indeed, in $S(\Phi)$. First, by (2.6)(c), there exists $\psi \in S$ whose stationary PSI ladder $\mathcal{S}(\psi) := (T_h)_h$ satisfies

$$\operatorname{dist}(f, T_h) \leq \operatorname{dist}(f, S_h) + o(h^{k-1}).$$

Therefore,

$$\operatorname{dist}(f, T_h) = o(h^{k-1}), \quad h = h_1, h_2, \ldots, \tag{4.5}$$

and hence ψ satisfies the required condition (ii).

The proof that ψ satisfies condition (i) is as follows. First, we define the bracket product $[f, g]$ of the L_2-functions f and g as follows:

$$[f, g] := \sum_{j \in 2\pi\mathbb{Z}^d} \widehat{f}(\cdot + j)\overline{\widehat{g}(\cdot + j)}.$$

The *Gramian* of Φ is then defined to be the $\Phi \times \Phi$ matrix whose (ϕ, ϕ') entry is $[\widehat{\phi}, \widehat{\phi}']$. It is proved in [BDR3] that under assumption (c) of Corollary 1.7, G is k times continuously differentiable and $G(0)$ is nonsingular. Thus G^{-1} exists on a neighborhood B of the origin and is also smooth there.

Now with τ the 2π-periodization of the restriction of $\widehat{\Phi}$ to B, in [BDR3], $[\widehat{\psi}, \widehat{\psi}]$ is identified (around the origin) as $\tau^* G^{-1} \tau$. Since $\widehat{\Phi} \subset C^k$ (by virtue of the decay assumptions on Φ), it follows that $[\widehat{\psi}, \widehat{\psi}] \in C^k(B)$. In addition, in [BDR3], $\widehat{\psi}$ is computed as $G^{-1}\widehat{\Phi}$, which again shows that $\widehat{\psi} \in C^k(B)$. Combining all these observations, we readily conclude that $\Lambda_\psi^2 \in C^k(B)$ as well. □

Proof of Proposition 4.3. We follow the proof of Proposition 4.2 to obtain the "superfunction" ψ as detailed in that proof. As in that proof, we need only show that $\Lambda_\psi^2 \in C^k(B)$.

We argue the smoothness of Λ_ψ^2 as follows. First, since we know that $S(\Phi)$ is generated by compactly supported functions (viz, Φ), (2.6)(b) allows us to assume that $S(\psi)$ is generated by a compactly supported function, which we may assume without loss to be ψ itself. Furthermore, since we are in a univariate situation, we may invoke of [R1, Theorem 1.1] and [R2, Theorem 3.7]: combined, these results say that every univariate PSI space that is generated by a compactly supported L_2-function ψ is also generated by a compactly supported L_2-function whose Fourier transform does not have a 2π-periodic zero. Again, we may assume without loss of generality that our generator ψ is already the "favorable" generator of [R1] and [R2]. This implies that the denominator in the definition of Λ_ψ vanishes nowhere, and a simple application of Poisson's summation formula then yields that Λ_ψ^2 is real analytic and, in particular, analytic around the origin. □

REFERENCES

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

[BDR1] C. de Boor, R. A. DeVore, and A. Ron, *Approximation from shift-invariant subspaces of $L_2(\mathbb{R}^d)$*, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806

[BDR2] C. de Boor, R. DeVore and A. Ron, *The structure of shift invariant spaces and applications to approximation theory*, Technical Report CMS-TSR 92–08, University of Wisconsin, Madison, WI, 1992; J. Funct. Anal., 119 (1994), pp. 37–78.

[BDR3] C. de Boor, R.A. DeVore, and A. Ron, *Approximation orders of FSI spaces*, Constr. Approx., 13 (1997), to appear.

[BHR] C. de Boor, K. Höllig, and S. D. Riemenschneider, *Box Splines*, Springer-Verlag, Berlin, 1992.

[BR] C. de Boor and A. Ron, *Fourier analysis of approximation power of principal shift-invariant spaces*, Constr. Approx., 8 (1992), pp. 427–462.

[CDM] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 453, AMS, Providence, RI, 1991.

[D] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[HSS] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., to appear.

[J] R.-Q. Jia, *The Toeplitz theorem and its applications to approximation theory and linear PDE's*, Trans. Amer. Math. Soc., 7 (1995), pp. 2585–2594.

[Jo1] M. J. Johnson, *An upper bound on the approximation power of principal shift-invariant spaces*, Constr. Approx., to appear.

[Jo2] M. J. Johnson, *On the approximation power of principal shift-invariant subspaces of $L_p(\mathbb{R}^d)$*, J. Approx. Theory, to appear.

[JM] R.-Q. Jia and C.A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* II: *Powers of two*, Curves and Surfaces, P. J. Laurent, A. Le Méhaué, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[M] Y. Meyer, *Ondelettes et Opérateurs* I: *Ondelettes*, Hermann, Paris, 1990.

[R1] A. Ron, *A necessary and sufficient condition for the linear independence of the integer translates of a compactly supported distribution*, Constr. Approx., 5 (1989), pp. 297–308.

[R2] A. Ron, *Factorization theorems of univariate splines on regular grids*, Israel J. Math., 70 (1990), pp. 48–68.

SIAM J. MATH. ANAL.
Vol. 28, No. 3, pp. 749–767, May 1997

ANALYTIC FUNCTIONS OPTIMIZING COMPETING CONSTRAINTS*

J. WILLIAM HELTON† AND ANDREI E. VITYAEV‡

Abstract. Optimization of sup-norm-type performance functions over the space of H^∞ functions is an area of extensive research. In electrical engineering, it is central to the subject of H^∞ design, while in several complex variables, it is often required to produce analytic discs with valuable properties.

It has been known for many years that an H^∞-type optimum is frequency independent (flat). In this paper, we study simultaneous (Pareto) optimization of several competing performances $\Gamma_1, \dots, \Gamma_l$.

We find under strong assumptions on the performance functions that if we are optimizing over N functions $(f_1, \dots, f_N)$ in H^∞ and have l performance measures with $l \leq N$, then at a nondegenerate Pareto optimum $(f_1^*, \dots, f_N^*)$, *every* performance is flat.

Besides flatness, there are other *gradient–alignment* conditions which must hold at an optimum. The article presents these and thus gives the precise *first-derivative* test for a natural class of H^∞ Pareto optima.

Such optimality conditions are valuable for assessing how iterations in a computer run are progressing. Also, in the traditional case, optimality conditions have been the base of highly sucessful computer algorithms; see [J. W. Helton, O. Merino, and T. Walker, Indiana U. Math. J., 42 (1993), pp. 839–874].

Key words. H^∞ control, frequency response methods, analytic discs

AMS subject classification. 93C80

PII. S0036141095293086

1. Introduction.

This paper analyzes a problem in which one optimizes objective functions over the space H^∞_N of vector-valued functions $f = (f_1, \dots, f_N)$ defined on the unit circle, $\mathbf{T}$, where each coordinate function f_j belongs to $L^\infty(\mathbf{T})$ and extends to be analytic on the entire unit disk.

The objectives that we optimize are described in terms of nonnegative continuous functions Γ defined on $\mathbf{T} \times \mathbf{C}^N$. Given positive functions $\Gamma_j(e^{i\theta}, z) \in C^1(\mathbf{T}, \mathbf{C}^N)$, $j = 1, \dots, l$, and a function $f \in H^\infty_N$, we define the l performances

$$\gamma_j(f) := \sup_{\theta \in \mathbf{T}} \Gamma_j(e^{i\theta}, f), \quad j = 1, \dots, l.$$

The goals of this paper are best illustrated by restricting our study to the case of two performance functions Γ_1 and Γ_2, even though our results hold for l-performance functions.

Definition. *A function $f^* \in H^\infty_N$ is called a* Pareto optimum *for Γ_1 and Γ_2 if for each $f \in H^\infty_N$ one of the following two inequalities holds:*

$$\gamma_1(f) \geq \gamma_1(f^*) \quad \textit{or} \quad \gamma_2(f) \geq \gamma_2(f^*).$$

* Received by the editors October 11, 1995; accepted for publication January 9, 1996. An announcement of these results was presented at the 34th Conference on Decision and Control, New Orleans, LA, December 13–15, 1996.

http://www.siam.org/journals/sima/28-3/29308.html

† Department of Mathematics, University of California at San Diego, San Diego, CA 92093-0112 (helton@osiris.ucsd.edu). The research of this author was partially supported by the Air Force Office of Scientific Research and the National Science Foundation.

‡ Department of Mathematics, University of California at San Diego, San Diego, CA 92093 (vityaev@euclid.ucsd.edu). The research of this author was partially supported by the National Science Foundation.

The book of Boyd and Barratt [BB] gives a good discussion of Pareto optimality.

1.1. Degenerate versus nondegenerate Pareto optima. A function f^* can be a Pareto optimum for Γ_1 and Γ_2 in two basic ways.

Degenerate optima. The first case is where f^* can optimize Γ_1, that is,

$$\gamma_1(f^*+h) \geq \gamma_1(f^*) \quad \forall h \in H_N^\infty, \quad h \neq 0,$$

in which case Γ_2 is irrelevant. Similarily, f^* can optimize Γ_2, and then Γ_1 is irrelevant. This case has been seriously studied, and the main optimality result is stated in Theorem 1.1 below.

Nondegenerate optima. The second case is when both Γ_1 and Γ_2 are relevant. In this case, there is a pair of analytic functions h_1 and h_2 such that

$$\gamma_1(f^*+h_1) < \gamma_1(f^*),$$

$$\gamma_2(f^*+h_2) < \gamma_2(f^*).$$

In other words, we can improve each performance *separately* by adding h_1 or h_2 to f^*, but we cannot improve both performances at the same time.

The case of nondegenerate optima is the subject of this paper.

Example for $N=2$.

$$\begin{aligned}\Gamma_1(e^{i\theta},z) &= |\psi_1(e^{i\theta})-z_1|^2,\\ \Gamma_2(e^{i\theta},z) &= |\psi_2(e^{i\theta})-z_2|^2,\end{aligned}$$

where ψ_j are rational. In this case, the problem is separable into two completely independent one-dimensional single-performance problems. Therefore, there exists no nondegenerate Pareto optimum for Γ_1 and Γ_2. An example of a degenerate Pareto optimum would be any pair of functions $f=(f_1,f_2)$ such that either f_1 or f_2 is an optimum for Γ_1 alone or Γ_2 alone, respectively.

If, on the other hand, we consider the problem with

$$\begin{aligned}\widetilde{\Gamma_1}(e^{i\theta},z) &= |\psi_1(e^{i\theta})-z_1|^2,\\ \widetilde{\Gamma_2}(e^{i\theta},z) &= |\psi_2(e^{i\theta})-z_1-z_2|^2,\end{aligned}$$

then the problem cannot be separated into two independent single-performance problems, and for generic ψ_1 and ψ_2, almost all Pareto optima are nondegenerate.

1.2. A characterization of Pareto optima. The main result of this paper is that for a special class of Γ_j's ,namely the ones that are the norms of certain rational functions, a nondegenerate local pareto optimum $f^* \in H_N^\infty$ for $N>1$ satisfies

$$\Gamma_1(e^{i\theta}, f^*(e^{i\theta})) = \text{const.}_1$$

and

$$\Gamma_2(e^{i\theta}, f^*(e^{i\theta})) = \text{const.}_2$$

for all θ. *The striking fact is that both performances are flat.* See Theorem 2.1 for the precise statement of this result in the general case of l-performance functions

Note that if $N=1$ then both performances are almost never flat. Also, we give a result that indicates that there is a large class of Γ_j's for which flatness will not hold.

An earlier instance of the flatness condition (Theorem 2.1(I)) for Pareto optima was discovered by Young (see [PY] for proofs). It applies to jointly minimizing the first, second, third, etc. singular values of matrix-valued functions, which is quite a different context from the one in this paper.

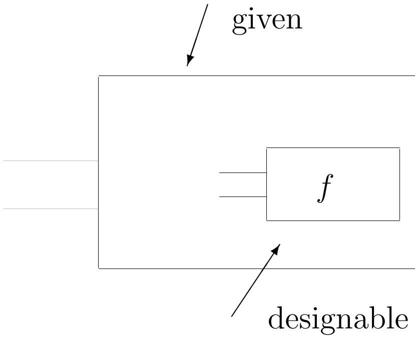


FIG. 1. *For a given plant, we want to find the best designable part, represented by f.*

1.3. Engineering motivation. This type of problem is central to frequency-domain-system design problems, where stability is a key constraint. In particular, it is important to the area of H^∞ control [H], [Fr]. The basic physical idea is simple. Recall that a linear-time-invariant system has a frequency-response function F defined on the imaginary axis and that the system is stable if F has no poles in the closed right half-plane (RHP). The behavior of the system when excited with a pure sine wave of frequency ω is determined by $F(i\omega)$. The following often occurs in a design procedure. We are required to build a system S, but part of the system is given (we are stuck with it) and part of the system is designable (denote its frequency-response function by f); see Figure 1. The performance of the system S at frequency ω is a function $\Gamma(\omega, f(i\omega))$ which depends on ω and on one's choice of the designable subsystem f. Let us adopt the convention that large Γ is bad while small Γ is good. Then in a worst-case "broadband" design, we consider the worst performance over all frequencies

$$\sup_{\omega} \Gamma(\omega, f(i\omega))$$

and try to minimize it over all admissible f. If our main constraint is that the designable subsystem f must be stable, then the design problem becomes the problem of finding a Pareto optimum with one Γ after transforming the RHP to the unit disk. When $N > 1$, these problems usually pose serious difficulties since traditional graphical trial and error methods are inadequate.

A number of authors (Mayne, Polak, and Salucidean; Fan and Tits; Streit; Boyd; Daleh; Pearson; Doyle, Glover, and Packard; Helton, Merino, and Walker; and Sideris) have written computer programs to search for an optimal f^* with certain kinds of Γ; see, e.g., [BB], [D], [FKTW], [HMW], [MNPW], [Si], and [St].

1.4. The classical case: One performance function. If we restrict our attention to the degenerate Pareto optimum, then we end up with a classical case of optimizing a single-performance function.

Now we state an earlier result for a single-performance function which this paper extends. Let $\Gamma(e^{i\theta}, z)$ be continuous nonnegative function. We are trying to find $f^* \in H^\infty_N$ which minimizes the following quantity:

$$\sup_{\theta \in \mathbf{R}} \Gamma(e^{i\theta}, f^*(e^{i\theta})).$$

Theorem 1.1 (see [H1]). *Let Γ be of class C^3 and $f^* \in H^\infty_N \cap C(\mathbf{T})$ be such that $(\partial\Gamma/\partial z)(e^{i\theta}, f^*(e^{i\theta}))$ never equals 0 on $\mathbf{T}$. If f^* is a local minimizer, then*

(I) *the function $\Gamma(\cdot, f^*(\cdot))$ is a constant on $\mathbf{T}$ and*

(II) *there exist functions $F \in H^1_N$ and $\lambda : \mathbf{T} \to \mathbf{R}^+$ measurable and positive almost everywhere on $\mathbf{T}$ such that*

$$\lambda(e^{i\theta})\frac{\partial\Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta})) = e^{i\theta}F(e^{i\theta}) \quad \text{a.e. on } \mathbf{T}.$$

When one adds a condition (III) asserting that Γ is convex in some directions, then one obtains a necessary and sufficient condition [HM1]. Indeed, [HM1] is the best reference for this result. Theorem 1.1 is extremely useful in that conditions (I) and (II) are a basis for computer diagnostics and software developed by Helton, Merino, and Walker; see [HMW].

1.5. Smooth performances versus multiple performances. In engineering applications, a single-performance function Γ can be used to incorporate several performance criteria. In particular, a Pareto optimum f^* for performance functions $\Gamma_1, \ldots, \Gamma_l$ can be viewed as a solution to the optimization problem with only one performance function defined by

$$\widetilde{\Gamma}(e^{i\theta}, z) := \max\left\{\frac{\Gamma_1(e^{i\theta}, z)}{\gamma_1(f^*)}, \ldots, \frac{\Gamma_l(e^{i\theta}, z)}{\gamma_l(f^*)}\right\}. \tag{1}$$

Namely, it is easy to check that f^* is a Pareto optimum for $\Gamma_1, \ldots, \Gamma_l$ if it is an optimum for $\widetilde{\Gamma}$. The main disadvantage of introducing $\widetilde{\Gamma}$ is that it is almost never differentiable, even though the Γ_j's are. As a consequence, the results proved for a single-performance-function optimization, reproduced in the theorem above, cannot be applied to $\widetilde{\Gamma}$.

1.6. Outline of the paper. This paper has the following structure. In section 2, we state and prove the main result of this paper. In section 3, we give two auxiliary results on uniqueness and existence of the Pareto optimum. In section 4, we reproduce the proofs of several lemmas which were proved by Trepreau in [T] and which have not been published. Section 4 is completely independent of the rest of the paper. In section 5, we discuss the connection between the M-OPT problem which had its origin in engineering mathematics and the analytic-disc techniques used in the popular several-complex-variables problem of extending a function defined on a manifold M in $\mathbf{C}^N$ to a function analytic in a neighborhood of a given point on M.

2. First-order conditions.

In this section, we give a precise statement of our results in the general case of l-performance functions.

Definition. *A function $f^* \in H^\infty_N$ is called a* local Pareto optimum *for $\Gamma_1, \ldots, \Gamma_l$ if there exists $\epsilon > 0$ such that*

$$\textit{for all } f \in H^\infty_N \textit{ and } \|f - f^*\| < \epsilon, \quad \textit{there exists } j, \quad \gamma_j(f) \geq \gamma_j(f^*).$$

For $l = 1$, this definition means that f^* minimizes $\sup_\theta \Gamma(e^{i\theta}, f(e^{i\theta}))$.

2.1. Main results. We introduce the notation

$$\frac{\partial\Gamma}{\partial z} = \begin{pmatrix} \frac{\partial}{\partial z_1}\Gamma_1 & \cdots & \frac{\partial}{\partial z_N}\Gamma_1 \\ \cdots & \cdots & \cdots \\ \frac{\partial}{\partial z_1}\Gamma_l & \cdots & \frac{\partial}{\partial z_N}\Gamma_l \end{pmatrix}. \tag{2}$$

Denote the unit disc in $\mathbf{C}$ by Δ.

We will impose the following assumption on performance functions.

ASSUMPTION 1. *Suppose $N \geq l$ and suppose $\Gamma_j = |P_j(e^{i\theta}, z)/Q_j(e^{i\theta}, z)|^2$, where P_j and Q_j are holomorphic polynomials in z with coefficients which are rational functions in $e^{i\theta}$. We assume, in addition, that the coefficients do not have poles on $\mathbf{T}$.*

We will consider a candidate for Pareto optimum f^* for which the following assumption holds.

ASSUMPTION 2. *The function $f^* \in H^\infty_N \cap C^\alpha$, $\alpha > 1/2$, with performances $\gamma_1^*, \dots, \gamma_l^*$ satisfies the following condition:*

There exists an analytic direction $h_j \in H^\infty_N \cap C^\alpha$, $\alpha > 1/2$, that improves all performances except for γ_j^*. Namely, there exist $C > 0$ and $t_0 > 0$ such that for every $t < t_0$,

$$\sup_\theta \Gamma_k(e^{i\theta}, f^*(e^{i\theta}) + th_j(e^{i\theta})) - \sup_\theta \Gamma_k(e^{i\theta}, f^*(e^{i\theta})) < -Ct \quad \textit{for } k = 1, \dots, l, \quad k \neq j.$$

Assumption 2 means that f^* is a nondegenerate Pareto optimum as discussed earlier, i.e., all l performances $\Gamma_1, \dots \Gamma_l$ play an active role. If not all of them matter, we have a smaller l.

Now we state the main result of this paper.

THEOREM 2.1. *Suppose the performance functions Γ_j satisfy Assumption* 1. *Suppose that $f^* \in H^\infty_N \cap C^\alpha$ is a local Pareto optimum that satisfies Assumption* 2. *Suppose further that $\frac{\partial \Gamma_j}{\partial z}(e^{i\theta}, f^*(e^{i\theta})) \in C^\alpha$ with $\alpha > 1/2$ and that*

$$\operatorname{rank} \frac{\partial \Gamma(e^{i\theta}, f^*(e^{i\theta}))}{\partial z} = l \tag{3}$$

for every $e^{i\theta} \in \mathbf{T}$.

Then the following hold:

(I) Flatness:

$$\Gamma_j(e^{i\theta}, f^*(e^{i\theta})) = \text{const.}, \quad j = 1, \dots, l.$$

(II) Gradient alignment: *There exists a row-vector-valued function $\lambda \in C^\alpha_l(\mathbf{T}, \mathbf{R})$, $\lambda \not\equiv 0$, with nonnegative entries, such that*

$$\lambda(e^{i\theta}) \frac{\partial \Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta})) = e^{i\theta} F, \quad F \in H^2_N.$$

Here $\frac{\partial \Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta}))$ denotes the $l \times N$ derivative matrix (2) *evaluated at $z = f(e^{i\theta})$.*

Remark. We will, in fact, show that for (II) to hold, it is enough to assume only that f^* is an optimum such that $f^* \in C^\alpha$ with $\alpha > 1/2$ and the rank condition (3) holds. In other words, Assumptions 1 and 2 are not needed for (II).

2.2. The classical Riemann–Hilbert problem. The main step in the proof of the flatness condition is solving the following version of Riemann–Hilbert problem:

Given an $l \times N$ matrix-valued function $A \in C^\alpha(\mathbf{T})$ with invertible values, given a closed interval $I \subset \mathbf{T}$, $I \neq \mathbf{T}$, find $h \in H^\infty_N$ such that

$$\begin{aligned} (\operatorname{Re}(Ah))_1 &> 0 \quad \text{for } e^{i\theta} \in I, \\ (\operatorname{Re}(Ah))_j &> 0 \quad \text{for all } e^{i\theta}, \quad j = 2, \dots, l. \end{aligned} \tag{4}$$

Here $(\cdot)_j$ stands for taking the jth entry of a vector. (Later the matrix A will be taken to be equal to $\frac{\partial \Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta}))$).

(4) is, in fact, a problem about the range of the Riemann–Hilbert map

$$w \longrightarrow 2\mathrm{Re}\,(Aw).$$

Questions about the range of the Riemann–Hilbert map arise in several aspects. For example, the gradient-alignment condition (Theorem 2.1(II)) will be shown to be equivalent to the fact that the range of the derivative map

$$w \longrightarrow 2\mathrm{Re}\left(\frac{\partial\Gamma(e^{i\theta}, f^*(e^{i\theta}))}{\partial z} w\right)$$

does not contain strictly positive functions, i.e., that all performances cannot be improved to first order at the same time.

Questions about the range of the Riemann–Hilbert map also arise in the analytic-disc techniques in one theoretical complex-variables problem; see section 6 for more details. See [Ve] as a standard reference on the theory of the Riemann–Hilbert problem.

In this paper, we give conditions on A which insure that problem (4) always has a solution.

To state our main condition, we need the following definition.

Definition. *A function $u(e^{i\theta}) \in L^\infty(\mathbf{T}, \mathbf{C})$ has a* pseudomeromorphic continuation *inside Δ if there exists a function $\widetilde{u}$, meromorphic in Δ and with finitely many poles in Δ, such that*

$$\lim_{r\to 1} \widetilde{u}(re^{i\theta}) = u(e^{i\theta}) \quad \text{a.e. } \mathbf{T}.$$

Theorem 2.2. *Suppose that $l \le N$. Suppose that an $l \times N$ matrix-valued function $A \in C^\alpha(\mathbf{T}, \mathbf{C})$, $\alpha > 1/2$, takes values of rank l on $\mathbf{T}$. If*

(i) *the entries of A have pseudomeromorphic continuation inside Δ*
or, more generally,

(ii) *the matrix A can be written as $D\widetilde{A}$, where D is an invertible diagonal matrix function and $\widetilde{A}$ has a pseudomeromorphic continuation inside Δ,*
then problem (4) *has a solution h.*

3. Proofs.

3.1. Proof of the flatness condition (Theorem 2.1(I)). First, we observe that every Γ_j has a Taylor expansion

$$\Gamma_j(e^{i\theta}, z + tw) = \Gamma_j(e^{i\theta}, z) + t2\mathrm{Re}\left(\frac{\partial\Gamma_j}{\partial z}(e^{i\theta}, z)\cdot w\right) + O(t^2). \tag{5}$$

Now suppose that $f^* \in H_N^\infty \cap C^\alpha$ is a minimizer which produces performances $\gamma_1^*, \dots, \gamma_l^*$ and that the performance function $\Gamma_1(e^{i\theta}, f^*(e^{i\theta}))$ is not constant. Then we can find $\epsilon_0 > 0$ and $I \subset \mathbf{T}$ so that $\mathbf{T} \setminus I$ is an open, nonempty interval and

$$\Gamma_1(e^{i\theta}, f^*(e^{i\theta}))|_{\mathbf{T}\setminus I} \le \gamma_1^* - \epsilon_0 = \sup_\theta \Gamma_1(e^{i\theta}, f^*(e^{i\theta})) - \epsilon_0. \tag{6}$$

We first want to find a vector-valued function $w \in H_N^\infty \cap C^\alpha$ that satisfies

$$\begin{pmatrix} \dfrac{\partial\Gamma_1}{\partial z_1}(e^{i\theta}, f^*(e^{i\theta})) & \cdots & \dfrac{\partial\Gamma_1}{\partial z_N}(e^{i\theta}, f^*(e^{i\theta})) \\ \cdots & \cdots & \cdots \\ \dfrac{\partial\Gamma_l}{\partial z_1}(e^{i\theta}, f^*(e^{i\theta})) & \cdots & \dfrac{\partial\Gamma_l}{\partial z_N}(e^{i\theta}, f^*(e^{i\theta})) \end{pmatrix} \begin{pmatrix} w_1 \\ \cdots \\ w_N \end{pmatrix} = \begin{pmatrix} \varphi \\ 0 \\ \cdots \\ 0 \end{pmatrix}, \tag{7}$$

where φ is an arbitrary function that never equals 0 on I.

Recall that $\Gamma_j(e^{i\theta}, z) = (P_j/Q_j)\overline{(P_j/Q_j)}$, where $P_j(e^{i\theta}, z)$ and $Q(e^{i\theta}, z)$ are holomorphic polynomials in z with rational coefficients depending on $e^{i\theta}$. The derivative is given by

$$\begin{aligned}\frac{\partial \Gamma_j}{\partial z_k}&(e^{i\theta}, f^*)\\ &= \frac{\dfrac{\partial P_j}{\partial z_k}(e^{i\theta}, f^*)Q_j(e^{i\theta}, f^*) - P_j(e^{i\theta}, f^*)\dfrac{\partial Q_j}{\partial z_k}(e^{i\theta}, f^*)}{[Q_j(e^{i\theta}, f^*)]^2}\overline{P_j(e^{i\theta}, f^*)Q_j(e^{i\theta}, f^*)}^{-1}.\end{aligned}$$

We introduce the notation

$$\Psi_j(e^{i\theta}) := \overline{P_j(e^{i\theta}, f^*(e^{i\theta}))}^{-1}\overline{Q_j(e^{i\theta}, f^*(e^{i\theta}))}[Q_j(e^{i\theta}, f^*(e^{i\theta}))]^2.$$

Note that $P_j \neq 0$ because $\frac{\partial \Gamma}{\partial z}$ has maximal rank. Then $\frac{\partial \Gamma_j}{\partial z_k}(e^{i\theta}, f^*(e^{i\theta}))\Psi_j(e^{i\theta})$ can be extended meromorphically inside the unit disc. By Assumption 1, the meromorphic functions $\frac{\partial \Gamma_j}{\partial z_k}(z, f^*(z))\Psi_j(z)$ do not have poles on the boundary of the unit disc or accumulating to the boundary. Let $\beta(z)$ be the finite Blaschke product such that $\beta(z)\frac{\partial \Gamma_j}{\partial z_k}(z, f^*(z))\Psi_j(z)$ is holomorphic in the unit disc for $j = 1, \dots, l$ and $k = 1, \dots N$.

Multiply both sides of (7) on the left by the $l \times l$ diagonal matrix D, which has the diagonal entries $\Psi_1(e^{i\theta}), \dots, \Psi_l(e^{i\theta})$. Note that each $\Psi_j(e^{i\theta})$ does not vanish anywhere on $\mathbf{T}$.

By our assumptions, the matrix $D\frac{\partial \Gamma}{\partial z}$ has rank l everywhere. Since it is also of class C^α with $\alpha > 1/2$, by Proposition 5.1, there exists the $N \times N$ constant matrix H such that the first l columns of the product $D\frac{\partial \Gamma}{\partial z}H$ are linearly independent everywhere on $\mathbf{T}$.

Denote by B the the first l columns of the holomorphic matrix $D\frac{\partial \Gamma}{\partial z}H\beta$. Let $\widetilde{B}$ be an $l \times l$ holomorphic matrix-valued function such that

$$B\widetilde{B} = (\det B)I_{l\times l}.$$

Then the vector

$$\begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_N \end{pmatrix} = \beta H \begin{pmatrix} \widetilde{B} \\ 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{pmatrix} \tag{8}$$

satisfies (7) with $\varphi(e^{i\theta}) \neq 0$ for all θ. Here $\binom{\widetilde{B}}{0}$ is an $N \times l$ matrix with the last $N - l$ rows equal to zero.

Since φ is nonzero on I, its argument $\arg\varphi|_I$ is well defined. Extend $\varphi|_I$ to the whole $\mathbf{T}$ in such a way that the extension $\widetilde{\varphi}$ has winding number zero. Let h be a holomorphic function such that $\arg h = -\arg\widetilde{\varphi}$. Then $h\varphi|_I$ is real valued and positive, and therefore the vector $-hw$ has the property that

$$2\mathrm{Re}\left(\frac{\partial}{\partial z}\Gamma_1(e^{i\theta}, f^*(e^{i\theta})) \cdot [-hw]\right) < -\epsilon_1, \quad e^{i\theta} \in I.$$

Choose ϵ_2 small enough so that

$$2\mathrm{Re}\left(\frac{\partial}{\partial z}\Gamma_1(e^{i\theta}, f^*(e^{i\theta})) \cdot [-\epsilon_2 hw]\right) \leq \epsilon_0/2, \quad e^{i\theta} \in \mathbf{T} \setminus I.$$

Then there exist positive constants C and t_0 such that for any $t \leq t_0$, the following holds:

$$\sup_\theta \left[\Gamma_1(e^{i\theta}, f^*(e^{i\theta})) - t2\mathrm{Re}\left(\frac{\partial}{\partial z}\Gamma_1(e^{i\theta}, f^*(e^{i\theta}))hw\right)\right] - \sup_\theta \Gamma_1(e^{i\theta}, f^*(e^{i\theta})) < -Ct,$$

$$2\mathrm{Re}\left(\frac{\partial}{\partial z}\Gamma_j(f^*)hw\right) = 0, \quad j = 2, \dots, l.$$

In other words, we have produced the analytic direction $-hw$ that "improves" the performance γ_1 and leaves the performances $\gamma_2, \dots, \gamma_l$ "unchanged" up to the first order.

By Assumption 2, there exists a direction $v \in H^\infty_N \cap C^\alpha$ that "improves" the performances $\gamma_2, \dots, \gamma_l$:

$$\sup_\theta \Gamma_j(e^{i\theta}, f^*(e^{i\theta}) + tv(e^{i\theta})) - \sup_\theta \Gamma_j(e^{i\theta}, f^*(e^{i\theta})) < -C't, \quad j = 2, \dots, l.$$

Now consider $\widetilde{w} = -hw + \varepsilon v$ for ε small enough. Then the analytic direction $\widetilde{w}$ "improves" all Γ_j's. Therefore, by (5), for small t, the function $f^* + t\widetilde{w}$ has better performances:

$$\gamma_j(f^* + t\widetilde{w}) < \gamma_j(f^*), \quad j = 1, \dots, l.$$

We have reached a contradiction. □

Proof of Theorem 2.2. The proof is a line-by-line repetition of a part of the proof of Theorem 2.1(I). □

3.2. Counterexample. Now we give some indications that we need Assumption 1 on Γ's for flatness to hold. The flatness result, if it holds, would imply that if $\Gamma_2(e^{i\theta}, f(e^{i\theta})), \dots, \Gamma_l(e^{i\theta}, f(e^{i\theta}))$ are constants and $\Gamma_1(e^{i\theta}, f(e^{i\theta}))$ satisfies (6), then we can find an analytic vector h such that

$$\begin{aligned} \mathrm{Re}\left(\tfrac{\partial\Gamma_1}{\partial z}(e^{i\theta}, f(e^{i\theta})) \cdot h\right) > 0 \quad &\text{for } e^{i\theta} \in I, \\ \mathrm{Re}\left(\tfrac{\partial\Gamma_j}{\partial z}(e^{i\theta}, f(e^{i\theta})) \cdot h\right) > 0 \quad &\text{for all } e^{i\theta}, \quad j = 2, \dots, l. \end{aligned}$$

Considering the question of existence of such an h in a little more general setting leads to problem (4).

Proposition 3.2 below shows that there exists an A such that problem (4) is not solvable. While we have not done so, we suspect that one can construct a simple A for which (4) is not solvable and which can be written as $\frac{\partial\Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta}))$ for some Γ_j's and f^*.

We start with the following lemma which gives a characterization of the derivative map (see [BRT] and [T]).

Lemma 3.1. *Assume $N \geq l$ and suppose that A is an $l \times N$ complex matrix-valued function on $\mathbf{T}$ of class C^α with $1/2 < \alpha < 1$ and which has maximal rank ℓ at every point on $\mathbf{T}$. Consider the following Riemann–Hilbert operator:*

$$\mathcal{F} : H^2_N \longrightarrow L^2_l(\mathbf{T}, \mathbf{R}), \qquad \mathcal{F}(w) = 2\mathrm{Re}(Aw).$$

Then $(\mathrm{range}\mathcal{F})^\perp \subset L^2(\mathbf{T}, \mathbf{R})$ *consists of all such real-valued ℓ-vectors $g \in L^2$ so that*

$$A^t g \in zH^2_N. \tag{9}$$

Proof. Suppose that $g \in L^2$ belongs to $(\mathrm{range}\mathcal{F})^\perp$. Then

$$\int g \cdot \mathrm{Re} Aw = 0 \quad \forall w \in H^2. \tag{10}$$

Since this is true for iw, we have

$$\int g \cdot Aw = 0 \quad \forall w \in H^2. \tag{11}$$

Therefore, $(\overline{A}^\top g, w) = 0$, where $(\cdot,\cdot)$ is the inner product in $L^2(\mathbf{T},\mathbf{C})$. Then (9) follows. □

Remark. If we denote the restriction of $\mathcal{F}$ to $H^2_N \cap C^\alpha$ by $\mathcal{F}^\alpha$, then the continuity of $\mathcal{F}$ implies that

$$(\mathrm{range}\mathcal{F}^\alpha)^\perp = (\mathrm{range}\mathcal{F})^\perp. \tag{12}$$

Proposition 3.2 below states that problem (4) cannot be solved for every A.

PROPOSITION 3.2. *Suppose that $N = l = 2$. Given any closed $I \subset \mathbf{T}$ with $I \neq \mathbf{T}$, there exists A such that for any $v \in C^\alpha(I)$ with $\alpha > 1/2$, $v > 0$,* on I *and any $\psi \in C^\alpha(\mathbf{T})$ with $\psi > 0$, there exists no solution $h \in H^\infty_2 \cap C^\alpha$ to the following Riemann–Hilbert problem:*

$$2\mathrm{Re}(Ah) = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}, \qquad \varphi|_I = v.$$

Proof. Take $g_0 = (p(e^{i\theta}), 1)$ with $p = 0$ on $\overline{\mathbf{T}\backslash I}$ and $p > 0$ on $\mathrm{int}(I)$. Then $(g_0, (\varphi,\psi))_{L^2} > 0$ for any $\psi > 0$ and any φ with $\varphi|_{\mathrm{int}(I)} > 0$.

Now we take

$$A = \begin{pmatrix} e^{i\theta} & 0 \\ -p(e^{i\theta})e^{i\theta} & e^{i\theta} \end{pmatrix}.$$

Obviously, $g_0 A \in zH^2$, and therefore g_0 belongs to $(\mathrm{range}\mathcal{F})^\perp$. □

3.3. Proof of Theorem 2.1(II). We first need the following theorem.

THEOREM 3.3. *Suppose that a subspace $\mathcal{R} \subset C^\alpha \subset L^2_l(\mathbf{T},\mathbf{R})$ has the property that its complement $(\mathcal{R})^\perp$ is finite dimensional and a subset of C^α_l. Suppose also that its closure in the L^2 topology $\overline{\mathcal{R}}$ satisfies $\overline{\mathcal{R}} \cap C^\alpha = \mathcal{R}$. If $\mathcal{R}$ does not contain vector functions with every component strictly positive, then there exists a function $\lambda \not\equiv 0$ in $(\mathcal{R})^\perp$ with every component positive (but not necessarily strictly positive).*

The proof requires the following.

LEMMA 3.4. *If the vectors $v_1, \ldots, v_m$ satisfy*

$$\{x : \exists j, 1 \leq j \leq m, \quad x \cdot v_j > 0\} = \mathbf{R}^n \backslash \{0\},$$

then the set

$$\mathrm{hull}(v_1, \ldots, v_m) := \{t_1 v_1 + \cdots + t_m v_m : \ t_j \geq 0\}$$

is the whole of $\mathbf{R}^n$

Proof. Abbreviate $\mathrm{hull}(v_1, \ldots, v_m)$ by V. Suppose that $V \neq \mathbf{R}^n$. Take any point x_0 which is not in V. Then there exists a plane $\{x : x \cdot \xi = r\}$ which separates x_0 and V, i.e., $x_0 \cdot \xi > r$ and $x \cdot \xi < r$ for every $x \in V$.

Since V is a cone, the latter implies that $x \cdot \xi \leq 0$ for every $x \in V$. In particular, it means that

$$v_j \cdot \xi \leq 0, \quad j = 1, \ldots, m.$$

We have reached a contradiction. $\square$

Proof of Theorem 3.3. Suppose that the functions $g_1, \ldots, g_n$ form a basis of $(\mathcal{R})^\perp$. We want to prove that there exists a linear combination $a_1 g_1(\zeta) + \cdots + a_n g_n(\zeta)$ in $(\mathcal{R})^\perp$ that is positive on all of $\mathbf{T}$. Here $a_j \in \mathbf{R}$. Denote the $n \times l$ matrix with rows $g_1(\zeta), \ldots, g_n(\zeta)$ by $g(\zeta)$. We claim that it is enough to show the inequality

$$\inf_{a \in \mathbf{R}^n, \|a\|=1} \max_{\zeta \in \mathbf{T}} a^t g(\zeta) \leq 0. \tag{13}$$

Here we use the notation

$$\max_{\zeta \in \mathbf{T}} (b_1(\zeta), \ldots, b_l(\zeta)) = \max(\max_{\zeta \in \mathbf{T}} b_1(\zeta), \ldots, \max_{\zeta \in \mathbf{T}} b_l(\zeta)).$$

Note that we do not take the absolute values of functions.

The quantity $\max_{\zeta \in \mathbf{T}} a^t g(\zeta)$ depends continuously on a, which varies over a compact set in $\mathbf{R}^n$. Therefore, if (13) holds, then there exists $a_0 \in \mathbf{R}^n$ such that

$$\max_{\zeta \in \mathbf{T}} a_0^t g(\zeta) \leq 0.$$

Hence all components of the vector function $-a_0^t g(\zeta)$ are positive and the claim follows.

Now we use a dual extension argument. This uses the fact that for any function $\mu \in C_l^\alpha(\mathbf{T})$,

$$\max_{\zeta \in \mathbf{T}} \mu(\zeta) = \sup_{u \in B_l^+} \int \mu(\zeta) \cdot u(\zeta) d\zeta,$$

where we use the notation $B_l^+ = \{u \in C_l^\alpha(\mathbf{T}, \mathbf{R}) : \|u\|_{L^1} = 1, u_j(\zeta) > 0,\ j = 1, \ldots, l\}$. Therefore, we need to prove the inequality

$$\inf_{a \in \mathbf{R}^n, \|a\|=1} \sup_{u \in B_l^+} \int (a^t g(\zeta)) \cdot u(\zeta) d\zeta \leq 0. \tag{14}$$

Given any $u \in B_l^+$, we define the vector $c(u) \in \mathbf{R}^n$ as follows:

$$c(u) := \left(\int g_1 \cdot u, \ldots, \int g_n \cdot u \right).$$

Lemma 3.5. *The set* $\{x : \exists u \in B_l^+, \ \ x \cdot c(u) > 0\}$ *is not the whole of* $\mathbf{R}^n \setminus \{0\}$.

Proof. Suppose the contrary: $\{x : \exists u \in B_l^+, \ \ x \cdot c(u) > 0\} = \mathbf{R}^n \setminus \{0\}$. Since, in fact, we have an open covering of the unit sphere S^{n-1}, which is compact, there exist $u_1, \ldots, u_m$ such that $\{x : \exists j, 1 \leq j \leq m, \ x \cdot c(u_j) > 0\} = \mathbf{R}^n \setminus \{0\}$. Then by Lemma 3.4, the convex hull $\operatorname{hull}(c(u_1), \ldots, c(u_m)) = \mathbf{R}^n$.

Therefore, there exists a positive, nontrivial linear combination of $c(u_j)$'s which is zero:

$$b \in \mathbf{R}^m, \quad b \neq 0, \quad b_j \geq 0, \quad \sum b_j c(u_j) = 0.$$

We consider $\widetilde{u}(\zeta) = \sum b_j u_j(\zeta)$. Then $\widetilde{u}(\zeta) \in C^\alpha$ would be a vector function with every component strictly positive. However, on the other hand,

$$\int g\widetilde{u} = \int \sum b_j(gu_j) = \sum b_j c(u_j) = 0,$$

which implies that $\widetilde{u} \in \overline{\mathcal{R}} \cap C^\alpha = \mathcal{R}$, i.e., we have reached a contradiction. □

Now we continue with the proof of Theorem 3.3. Take a point a_0 not in $\{x : \exists u \in B_l^+, \quad x \cdot c(u) > 0\}$ with $\|a_0\| = 1$. Then

$$\int (a_0^t g(\zeta)) \cdot u(\zeta) d\zeta = a_0^t \int g(\zeta) u(\zeta) d\zeta = a_0^t c(u) \leq 0 \quad \forall u \in B_l^+,$$

and therefore inequality (14) holds. □

Now we continue with the proof of Theorem 2.1(II). Since f^* is an optimum, there is no strictly positive function in the range of the derivative map

$$\mathcal{F}^\alpha : H_N^2 \cap C^\alpha \to C^\alpha, \qquad \mathcal{F}^\alpha(w)(\zeta) = 2\mathrm{Re}\left(\frac{\partial\Gamma(\zeta, f^*(\zeta))}{\partial z} w(\zeta)\right).$$

We observe that Lemma 5.5 and (12) imply that $(\mathrm{range}\mathcal{F}^\alpha)^\perp$ is a finite-dimensional subset of C^α. We claim that $\overline{\mathrm{range}\mathcal{F}^\alpha} \cap C^\alpha = \mathrm{range}\mathcal{F}^\alpha$. To show the claim, we first note that $\mathrm{range}\mathcal{F}$, where $\mathcal{F}$ is defined as in Lemma 3.1, is closed in L^2. Therefore, $\overline{\mathrm{range}\mathcal{F}^\alpha} = \mathrm{range}\mathcal{F}$. The C^α-hypoellipticity of $\mathcal{F}$ (see Lemma 5.5) implies that $\mathrm{range}\mathcal{F} \cap C^\alpha = \mathrm{range}\mathcal{F}^\alpha$, and the claim follows.

The argument above shows that we can apply Theorem 3.3 with $\mathcal{R} = \mathrm{range}\mathcal{F}^\alpha$. Therefore, there is a vector-valued function λ in $(\mathrm{range}\mathcal{F})^\perp$ with each component positive. By Lemma 3.1, λ must satisfy (9), which implies Theorem 2.1(II). □

4. Additional results.

4. Additional results. In this section, we state the results on uniqueness and existence of the M-OPT problem which are, in fact, easy corollaries of the results proved in [HMar], [HM1], and [V].

4.1. Uniqueness. Consider the sublevel sets

$$\mathcal{S}_\theta^j(\gamma) := \{z \in \mathbf{C}^N : \Gamma_j(e^{i\theta}, z) \leq \gamma\}.$$

THEOREM 4.1. *Suppose that the performance functions $\Gamma_1, \ldots, \Gamma_l$ satisfy Assumption 1. Suppose that the sublevel sets $\mathcal{S}_\theta^j(\gamma_j)$ are strictly convex for every $j = 1, \ldots, l$, every $\theta \in \mathbf{T}$, and every γ_j.*

Assume that the real numbers $\gamma_1^*, \ldots, \gamma_l^*$ have the property that for every function f with performances $\gamma_1(f), \ldots, \gamma_l(f)$ satisfying $\gamma_j(f) \leq \gamma_j^*$, Assumption 2 holds and the matrix $\partial\Gamma(e^{i\theta}, f(e^{i\theta}))/\partial z$ has rank l.

Then if $f^* \in H_N^\infty \cap C^\alpha$ with $\alpha > 1/2$ is a Pareto optimum with performances

$$\sup_\theta \Gamma_j(e^{i\theta}, f^*(e^{i\theta})) = \gamma_j(f^*),$$

this Pareto optimum is unique, namely, there is no other function $f \in H_N^\infty \cap C^\alpha$, $\alpha > 1/2$, with the property that $\gamma_j(f) = \gamma_j(f^*)$ for all $j = 1, \ldots, l$.

Proof. Suppose that such an f does exist. Then

$$f(e^{i\theta}) \in \partial\mathcal{S}_\theta^1(\gamma_1^*) \cap \cdots \cap \partial\mathcal{S}_\theta^l(\gamma_l^*) \quad \forall e^{i\theta} \in \mathbf{T} \tag{15}$$

by the flatness result. Here $\gamma_j^* = \gamma_j(f^*)$. Since (15) is true for f^* as well, the function $h = 1/2f + 1/2f^*$ has the property that

$$h(e^{i\theta}) \in \mathcal{S}_\theta^1(\gamma_1^*) \cap \cdots \cap \mathcal{S}_\theta^l(\gamma_l^*) \quad \forall e^{i\theta} \in \mathbf{T}.$$

Strict convexity implies that there exists θ_0 satisfying

$$h(e^{i\theta_0}) \in \mathrm{int}\mathcal{S}_{\theta_0}^1(\gamma_1^*) \cap \cdots \cap \mathrm{int}\mathcal{S}_{\theta_0}^l(\gamma_l^*). \tag{16}$$

If the performance functions evaluated at h are flat (i.e., the flatness result (Theorem 2.1(I)) holds for h), then (16) should hold for every θ, and therefore f^* is not an optimum. If the performance functions of h are not flat, then by Theorem 2.1, they all can be improved simultaneously, and therefore f^* is not an optimum in this case either. $\square$

4.2. Existence. The following theorem was proved in [HMar].

THEOREM 4.2. *Suppose that $\Gamma(e^{i\theta}, z)$ is a positive continuous function and suppose that the sublevel sets of Γ satisfy $\partial\mathcal{S}_\theta(\gamma) = \{z \in \mathbf{C}^N : \Gamma(e^{i\theta}, z) = \gamma\}$. Suppose further that the sublevel sets $\mathcal{S}_\theta(\gamma)$ are uniformly bounded and polynomially convex. Suppose that $f^n \in H_N^\infty$ satisfy $\lim_n \|\Gamma(e^{i\theta}, f^n(e^{i\theta}))\|_\infty = \gamma$. Let f be a normal limit of f^n. Then $\|\Gamma(e^{i\theta}, f(e^{i\theta}))\|_\infty \le \gamma$.*

As a corollary, we can state the following existence result.

THEOREM 4.3. *Suppose that $\Gamma_j, j = 1, \ldots, l$, are positive C^1 functions. Suppose that the sublevel sets satisfy $\partial\mathcal{S}_\theta^j(\gamma) = \{z \in \mathbf{C}^N : \Gamma_j(e^{i\theta}, z) = \gamma\}$ for $j = 1, \ldots, l$ and for every γ and θ. Suppose further that the sets $\mathcal{S}_\theta^1(\gamma) \cap \cdots \cap \mathcal{S}_\theta^l(\gamma)$ are uniformly bounded and polynomially convex for every γ.*

Then there exists a Pareto optimum for $\Gamma_1, \ldots, \Gamma_l$.

Remark. Note that we do not impose Assumption 1 in this theorem. In particular, the flatness property of the optimum needs not hold.

Proof. First, we reduce the problem to the case of one performance function by introducing

$$\widetilde{\Gamma}(e^{i\theta}, z) := \max\big(\Gamma_1(e^{i\theta}, z), \ldots, \Gamma_l(e^{i\theta}, z)\big).$$

We start with an initial guess $f^1 \equiv 1$. We then make further guesses, f^n's, improving $\widetilde{\Gamma}$ if possible. Since the $\widetilde{\mathcal{S}}_\theta(\gamma)$'s are uniformly bounded, the set $\{f^n\}$ is uniformly bounded in H_N^∞ and therefore has a subsequence, converging locally uniformly in the open unit disc to some limit f. Applying Theorem 4.2, we conclude that f is an optimum for $\widetilde{\Gamma}$ with $\sup_\theta \widetilde{\Gamma}(e^{i\theta}, f(e^{i\theta})) = \gamma$. $\square$

5. Technical lemmas. In this section, we reproduce the proofs of several lemmas that were proved by Trepreau and which have not been published. They can all be found in the preprint [T]. Lemma 5.5 is very close to the results proved in [BG].

PROPOSITION 5.1. *Suppose that $A \in C^\alpha(\mathbf{T})$ with $\alpha > 1/2$ is an $l \times N$ matrix-valued function of maximum rank l at every point on $\mathbf{T}$. Then there exists an $N \times N$ constant complex matrix H such that the first l columns of AH are linearly independent at every point on $\mathbf{T}$.*

To prove this proposition, we will need the following lemma, in which the role of the assumption $\alpha > 1/2$ becomes clear.

LEMMA 5.2. *Given $u(x) \in C^\alpha([0,1], \mathbf{C})$ with $\alpha > 1/2$,* the range *of u has zero Lebesgue measure in $\mathbf{C}$.*

Proof. Divide the interval [0,1] into n equal parts by points j/n, $j = 0, \ldots, n$. Then the range of u can be covered by the n discs of radius $C(1/n)^\alpha$ centered at $u(j/n)$, $j = 0, 1, \ldots, n$, where C is the Hölder constant of u. The total measure of these discs does not exceed

$$n\pi(C(1/n)^\alpha)^2 = C^2\pi n^{1-2\alpha}.$$

Since $\alpha > 1/2$, this quantity tends to zero as n approaches infinity. □

To prove Proposition 5.1, we begin by proving a special case.

LEMMA 5.3. *Let $a \in C^\alpha(\mathbf{T}, \mathbf{C})$, $\alpha > 1/2$, be the N-vector such that $a \neq 0$ anywhere on $\mathbf{T}$. Then there exists $\beta \in \mathbf{C}^N$ such that $a(e^{i\theta}) \cdot \beta$ vanishes nowhere on $\mathbf{T}$.*

Proof. We will prove the statement for any open subset I of $\mathbf{T}$ by induction on N. (We will assume that a is of class C^α *uniformly* on I.) Since the case where $N = 1$ is trivial, we assume that the lemma holds for $N - 1$. Let $Z \subset I$ be the zero set of a_N and $Z_0 \subset I$ be an open neighborhood of Z on which $(a_1, \ldots, a_{N-1}) \neq 0$. By the induction hypothesis, there exists $(\beta_1, \ldots, \beta_{N-1})$ such that $a_1\beta_1 + \cdots + a_{N-1}\beta_{N-1}$ does not vanish on Z_0. Also, the function $(a_1\beta_1 + \cdots + a_{N-1}\beta_{N-1})/a_N \in C^\alpha(I \setminus Z)$ cannot be onto $\mathbf{C}$ by Lemma 5.4 below. (Note that σ/a_N is bounded on $I \setminus Z_0$.)

Therefore, we can find a number $-\beta_N$ that is not in the range of σ/a_N, and so $a_1\beta_1 + \cdots + a_{N-1}\beta_{N-1} + a_N\beta_N$ vanishes nowhere on I. □

LEMMA 5.4. *Let $\sigma := a_1\beta_1 + \cdots + a_{N-1}\beta_{N-1}$. The range of the function $\sigma/a_N \in C^\alpha(Z_0 \setminus Z)$ is not the whole of $\mathbf{C}$.*

Proof. Note that $Z_0 \setminus Z$ is a countable union of open intervals (t_j, t_{j+1}). We will show that the range of each of the restrictions $\sigma/a_N \in C^\alpha((t_j, t_{j+1}))$ is of measure zero in $\mathbf{C}$. Note that $\sigma \neq 0$ on Z_0 and a_N can vanish only at the endpoints t_j. If it vanishes at t_j, then it satisfies

$$|a_N(t) - 0| \leq c|t - t_j|^\alpha, \tag{17}$$

which implies

$$|\sigma(t)/a_N(t)| \geq c'|t - t_j|^{-\alpha}. \tag{18}$$

Divide $\mathbf{C}$ into the annuli $L_k = \{z : k \leq |z| \leq k + 1\}$.

Away from points where a_N vanishes, we use the fact that both σ and a_N are of class C^α uniformly on $Z \setminus Z_0$ to conclude that the restriction of the range of $(\sigma/a_N)\mid_{(t_j,t_{j+1})}$ onto L_k is a curve which is *uniformly* C^α. In the neighborhood of points where a_N vanishes, we use estimate (18) to reach the same conclusion. Therefore, the range of $(\sigma/a_N)\mid_{(t_j,t_{j+1})}$ has zero measure in $\mathbf{C}$ by Lemma 5.2. □

Proof of Proposition 5.1. We write $A = (A_1, \ldots, A_N)$, where the A_j's are the columns of A. For $p = 1, \ldots, l$, we denote the vector formed by the first p components of A_j by A_j^p. By induction on p, we will show that there exists an invertible constant matrix H such that the matrix $((AH)_1^p, \ldots, (AH)_p^p)$ has rank p at every point.

Since A has rank l at every point, $(A_1^1, \ldots, A_N^1)$ has rank 1 and by Lemma 5.3 there exists $\langle\beta_1, \ldots, \beta_N\rangle$ such that $\beta_1A_1^1 + \cdots + \beta_NA_N^1$ does not vanish. Let H_1 be an invertible matrix with the first column $\langle\beta_1, \ldots, \beta_N\rangle^t$. Then $(AH_1)_1^1$ has rank 1 everywhere.

Now we assume that the claim is true for p. Replacing A by AH_p, we can assume that $(A_1^p, \ldots, A_p^p)$ has rank p everywhere. This implies that at every point on $\mathbf{T}$, one of the determinants

$$d_j = \det(A_1^{p+1}, \ldots, A_p^{p+1}, A_j^{p+1}), \quad j = p + 1, \ldots, N, \tag{19}$$

is nonzero. By Lemma 5.3, we can find $\beta_j, j = p+1, \dots, N$, such that $d_{p+1}\beta_{p+1} + \cdots + d_N\beta_N$ is nonzero on $\mathbf{T}$. Let $\widetilde{H}$ be a $(N-p)\times(N-p)$ invertible matrix whose first column is $\beta_{p+1}, \dots, \beta_N$ and set

$$H_{p+1} = \begin{pmatrix} I_{p\times p} & 0 \\ 0 & \widetilde{H} \end{pmatrix} \tag{20}$$

to obtain an N-dimensional matrix. Then the matrix whose columns are $(AH_{p+1})_1^{p+1}$, $\dots, (AH_{p+1})_{p+1}^{p+1}$ has the determinant equal to

$$\det\left(A_1^{p+1}, \dots, A_p^{p+1}, \sum_{j=p+1}^{N} \beta_j A_j^{p+1}\right) = d_{p+1}\beta_{p+1} + \cdots + d_N\beta_N$$

and therefore has rank $p+1$ everywhere. $\square$

Here is one more technical lemma that we used in the proofs.

LEMMA 5.5. *Let the map*

$$\mathcal{F}_A : H_N^2 \to L_l^2(\mathbf{T}, \mathbf{R})$$

be defined by $\mathcal{F}_A(w) = 2\mathrm{Re}(Aw)$, where the $l\times N$ matrix $A \in C^\alpha$ has rank l everywhere on $\mathbf{T}$. Then we have

$$(\mathrm{range}\mathcal{F}_A)^\perp \subset C^\alpha(\mathbf{T}, \mathbf{R}). \tag{21}$$

Proof. First, we prove (21) for the case where $N = l$. We will use the following result.

THEOREM 5.6 (see [Ve]). *Suppose that A is a square $l \times l$ matrix-valued function on $\mathbf{T}$ with invertible values such that $A \in C^\alpha(\mathbf{T})$. Then there exists a holomorphic matrix-valued function $S \in C^\alpha(\mathbf{T})$ such that $S^{-1} \in H_{l\times l}^2$ and such that*

$$\overline{A}^{-1}A = \overline{S}^{-1}DS. \tag{22}$$

Here D is the diagonal matrix with entries $e^{ik_1\theta}, \dots, e^{ik_N\theta}$, where k_j are the integers.

Vekua proved a slightly stronger result in his book [Ve, section 13, p. 97] by constructing the fundamental matrix of solutions for the Hilbert problem. However, he did not state his result in the form above since the Gohberg–Krein factorization, in which form the result is presented here, was discovered much later. The reduction is easy and can be found in, for example, [G].

To prove (21), consider the real-linear map

$$\mathcal{G} : q \in L^2(\mathbf{T}, \mathbf{R}) \to q + i\mathcal{H}(q) = w \to \mathrm{Re}(Aw) \in L^2(\mathbf{T}, \mathbf{R}). \tag{23}$$

Here $\mathcal{H}$ is the Hilbert transform with any particular fixed choice of normalization.

Suppose that $u \in L^2$ belongs to $(\mathrm{range}\mathcal{F}_A)^\perp$. Then $u \in \ker\mathcal{G}^\perp$, where $\mathcal{G}^\perp$ is the operator adjoint to $\mathcal{G}$ in $L^2(\mathbf{T}, \mathbf{R})$, i.e.,

$$\mathcal{G}^\perp(u) = \mathrm{Re}A^t u + \mathcal{H}(\mathrm{Im}A^t u). \tag{24}$$

Note that range$\mathcal{G}$ may be smaller than range$\mathcal{F}_A$ because of the normalization of $\mathcal{H}$. Therefore, $u \in L^2(\mathbf{T}, \mathbf{R})$ satisfies the following equation:

$$\mathrm{Re}A^t u + \mathcal{H}(\mathrm{Im}A^t u) = 0. \tag{25}$$

Let us introduce the holomorphic function $\psi \in H^2_N$:

$$\psi := \mathrm{Im} A^t u + i\mathcal{H}(\mathrm{Im} A^t u). \tag{26}$$

Then (25) and (26) imply that

$$\mathrm{Re} A^t u = -\mathrm{Im}\psi, \qquad \mathrm{Im} A^t u = \mathrm{Re}\psi. \tag{27}$$

Therefore, we have

$$A^t u = \mathrm{Re} A^t u + i\mathrm{Im} A^t u = -\mathrm{Im}\psi + i\mathrm{Re}\psi = i\psi, \tag{28}$$

$$\overline{A}^t u = \mathrm{Re} A^t u - i\mathrm{Im} A^t u = -\mathrm{Im}\psi - i\mathrm{Re}\psi = -i\overline{\psi}, \tag{29}$$

which can be rewritten as

$$u = (A^t)^{-1} i\psi, \tag{30}$$

$$u = -(\overline{A}^t)^{-1} i\overline{\psi}, \tag{31}$$

which implies

$$(A^t)^{-1} i\psi = -(\overline{A}^t)^{-1} i\overline{\psi}. \tag{32}$$

Thus ψ is a solution of the following Riemann–Hilbert problem:

$$\mathrm{Re}((A^t)^{-1}\psi) = 0. \tag{33}$$

By the results of Vekua, $\psi \in C^\alpha(\mathbf{T})$. Then (30) implies that the same is true for u. Inclusion (21) is proved for $N = l$.

For the case where $N > l$, we multiply the $l \times N$ matrix A by the $N \times N$ holomorphic matrix H from Proposition 5.1. Then we consider the operator $\mathcal{F}_{A,l}$ defined by the first l columns of AH. Thus we have

$$(\mathrm{range}\mathcal{F}_A)^\perp \subset (\mathrm{range}\mathcal{F}_{A,l})^\perp \subset C^\alpha(\mathbf{T}, \mathbf{R}). \qquad \square$$

6. Connection with the analytic-disc technique used in the problem of extending Cauchy–Riemann functions.

We discuss some connections between the M-OPT problem and the analytic-disc techniques used for studying the theoretical several-complex-variables problem of extending a function defined on a manifold M in $\mathbf{C}^N$ to a function analytic in a neighborhood of a given point on M.

6.1. Attached analytic discs correspond to the flatness condition of M-OPT.

First, we give some definitions. A smooth manifold $M = \{z \in \mathbf{C}^N : \rho_1(z) = \cdots = \rho_l(z) = 1\}$ is called *generic* if its defining functions $\rho_1, \ldots, \rho_l$ satisfy $\partial\rho_1 \wedge \cdots \wedge \partial\rho_l \neq 0$, where $\partial\rho_j$ are defined as in (2). We say that a continuous function u on M is a *CR function* if it is annihilated by all tangential antiholomorphic vector fields:

$$\sum_j a_j \frac{\partial}{\partial \bar{z}_j} u = 0 \quad \text{if} \quad \sum_j a_j \frac{\partial}{\partial \bar{z}_j} \rho_k = 0 \quad \forall k.$$

We study the question of local extendibility of CR functions on M holomorphically to some neighborhood of a given point p on M. The basic theorem for the analytic-disc method is the approximation theorem by Baouendi and Treves.

Theorem 6.1 (see [BT]). *Suppose that M is a generic manifold, p is a point on M, and u is a CR function on $M \cap \Omega$, where Ω is a neighborhood of p. Then there exists $\Omega' \subset \Omega$, a smaller neighborhood of p, such that u is a* (*uniform on $M \cap \Omega'$*) *limit of holomorphic polynomials.*

It follows from the maximum-modulus theorem that if we can fill some neighborhood of p with the interiors of analytic discs f attached to M (i.e., the boundary $f(\partial\Delta) \subset M$), then the function u can be extended holomorphically to (a part of) that neighborhood.

The construction of analytic discs attached to a manifold was the main tool used by many authors to prove the extendibility results; see [Tu1], [Tu2], and [BRT] as well as [Bog] and the references therein.

The main result of this paper on flatness (Theorem 2.1(I)), when stated geometrically, is an assertion about attached analytic discs. Namely, the flatness condition proved in [HM1] means that the optimal analytic disc f^* is attached to a loop of hypersurfaces $M_\theta = \{\Gamma(e^{i\theta}, z) = c\}$ in $\mathbf{C}^N$. The flatness condition proved in this paper says that f^* is attached to a loop of generic manifolds in $\mathbf{C}^N$, namely $M_\theta = \{\Gamma_j(e^{i\theta}, z) = c_j,\ j = 1, \ldots, l\}$.

6.2. The notion of defect versus the gradient-alignment condition. Since CR functions analytically continue to the set that is the union of "small" analytic discs attached to M, it is clear that this set is important. With this in mind, we let $\mathcal{R}_M$ denote the subset of $\mathbf{C}^N$ which is swept out by analytic discs attached to M. Be aware that not every CR function on M extends to $\mathcal{R}_M$ since the "small"-disc requirement has been dropped. Basic information about extension is provided by the notion of defect.

Definition. *Given an analytic disc f attached to a manifold $M = \{z \in \mathbf{C}^N : \rho_1(z) = \cdots = \rho_l(z) = 0\}$, we define the defect* def$f$ *of the disc f as the real dimension of the space*

$$\left\{\lambda \in C_l^\alpha(\mathbf{T}, \mathbf{R}) : \lambda(e^{i\theta}) \frac{\partial\rho(f(e^{i\theta}))}{\partial z} \in H_N^2\right\}.$$

It can be shown (see, e.g., [BRT]) that for any p on M, there exists $\epsilon(p)$ such that if $\|f\|_{C^\alpha} < \epsilon(p)$ then the defect is an integer between 1 and l.

The notion of defect was introduced for small discs by Tumanov in [Tu1]. The definition that we give was first introduced in [BRT]. The advantage of the latter is that it makes sense for any disc (not just small discs).

It was proved by Tumanov that if for any $\epsilon > 0$ there exists an analytic disc $f \in H_N^\infty \cap C^{1,\alpha}$ of defect zero and such that $\|f(e^{i\theta}) - p\|_{C^{1,\alpha}} < \epsilon$, then the interiors of small analytic discs attached to M sweep out a wedge with edge M near p, and therefore any CR function on M extends to that wedge.

With a given disc f attached to M with $f(0) = q$, we associate an M-OPT problem:

Find $h \in H_N^\infty$, which is a Pareto optimum for l-performance functions $\Gamma_{q1}, \ldots, \Gamma_{ql}$ defined by

$$\Gamma_{qj}(e^{i\theta}, z) = \rho_j(e^{i\theta} z + q), \quad j = 1, \ldots, l. \tag{34}$$

Note that

$$\frac{\partial\Gamma_q(e^{i\theta}, h(e^{i\theta}))}{\partial z} = e^{i\theta} \frac{\partial\rho(f(e^{i\theta}))}{\partial z}.$$

Then the condition that f is of defect zero is equivalent to

$$\lambda(e^{i\theta})\frac{\partial\Gamma(e^{i\theta},f(e^{i\theta}))}{\partial z}\notin e^{i\theta}H^2_N \quad \forall\lambda\in C^\alpha_l(\mathbf{T},\mathbf{R}), \tag{35}$$

which means that Theorem 2.1(II) is violated.

In other words, if disc f is of defect zero, then it cannot be a Pareto optimum for $\Gamma_{q1},\dots,\Gamma_{ql}$, where $q=f(0)$.

We use this observation to prove the following proposition, which shows that we can fill out an open set in $\mathbf{C}^N$ with discs attached to a hypersurface and close to a disc f of defect zero. Similar results of Tumanov require f to be small.

PROPOSITION 6.2. *Suppose that the disc $f\in H^\infty_N\cap C^\alpha$ is attached to a smooth hypersurface $M=\{z\in\mathbf{C}^N:\rho(z)=1\}$. Suppose that the defect of f is equal to zero. Then the set $\{g(0):g\in H^\infty_N, g(\partial\Delta)\subset M\}$ contains a neighborhood of $f(0)$ in $\mathbf{C}^N$.*

Note that we use H^∞ discs rather than $H^\infty\cap C^\alpha$ discs to fill a neighborhood in $\mathbf{C}^N$.

A similar result for the special case where M is a generic manifold of real codimension N can be found in [G]. The theorem in [G] is stated in terms of factorization indices of the matrix $\frac{\partial\rho(f)}{\partial z}$, but it can be restated as above.

Proof of Proposition 6.2. We can assume that $\rho(f(0))>1$. Consider the performance function Γ given by (34). Then (35) implies that Theorem 2.1(II) is violated and the quantity

$$\gamma(q)=\inf_{h\in H^\infty}\sup_\theta\Gamma_q(e^{i\theta},h(e^{i\theta}))$$

is less than 1.

We need the following lemma.

LEMMA 6.3. *The function $q\to\gamma(q)$ is continuous on $\mathbf{C}^N$.*

Proof. For any given ϵ, take $\delta>0$ such that

$$\|h_1-h_2\|_\infty\le\delta \implies \left\|\rho(h_1(e^{i\theta}))-\rho(h_2(e^{i\theta}))\right\|_\infty\le\epsilon.$$

Then by the definition of $\gamma(q)$, the inequality $\|q-\widetilde{q}\|<\delta$ implies $\gamma(\widetilde{q})\ge\gamma(q)-\epsilon$. At the same time, it implies $\gamma(q)\ge\gamma(\widetilde{q})-\epsilon$. Therefore, $\|\gamma(\widetilde{q})-\gamma(q)\|\le\epsilon$. □

Note that the function $q\to f^*_q$, where f^*_q is the optimum for Γ_q, needs not be continuous.

Now we continue with the proof of Proposition 6.2. By Lemma 6.3, there exists U, an open neighborhood of q in $\mathbf{C}^N$, such that $\gamma(\widetilde{q})<1$ for all $\widetilde{q}\in U$. Take a point $\widetilde{q}\in U$. Let $g\in H^\infty_N$ be a solution to the OPT problem with $\Gamma_{\widetilde{q}}$. Then $\rho(e^{i\theta}g(e^{i\theta})+\widetilde{q})<1$ for every θ. Denote by Ω the connected component of the set

$$\{z\in\Delta:\rho(g(z))>1\}$$

containing zero. Construct Ω^*, the set containing Ω, according to the following procedure. Take a point $\xi\in\Delta\setminus\Omega$. If there exists a closed Jordan curve in Ω which encircles ξ, then we set $\xi\in\Omega^*$. If there is no such curve, we set $\xi\notin\Omega^*$. Then Ω^* is an open simply connected region such that $\partial\Omega^*\subset\partial\Omega$, and therefore $g|_{\partial\Omega^*}\equiv 0$.

Let $\Psi:\Omega\to\Delta$ be the Riemann map such that $\Psi(0)=0$. Then $g\circ\Psi^{-1}$ is the analytic disc which is attached to M and which passes through $\widetilde{q}$. □

6.3. Manifolds of analytic discs. Motivated by the successful application mentioned above, the set of analytic discs attached to a manifold was studied in general. It was shown in [BRT] that the set of analytic discs, infinitely close to a point, forms an infinite-dimensional manifold. Namely, it was shown that the derivative map

$$\mathcal{F} : H^\infty_N \cap C^\alpha \to C^\alpha(\mathbf{T}, \mathbf{R}), \qquad \mathcal{F}(w) = 2\mathrm{Re}\left(\frac{\partial \rho(f(e^{i\theta}))}{\partial z} w\right) \tag{36}$$

is onto $C^\alpha(\mathbf{T}, \mathbf{R})$ for f close to a constant disc. Then application of the local-submersion theorem shows that we have a manifold.

In [F], [T], and [O], the restriction on the size of the disc f was removed and it was shown that under certain conditions the map (36) is onto.

In the engineering setup, if the set of discs close to f^* forms a manifold, then f^* is not an optimum. More precisely, suppose we are given a disc f^* attached to a loop of manifolds $M_\theta = \{\Gamma(e^{i\theta}, z) = c\}$. Then if the map (36) is onto, then there exists a holomorphic direction h such that

$$2\mathrm{Re}\left(\frac{\partial \Gamma}{\partial z} h\right) < 0,$$

and therefore $\gamma(f^* + th) < \gamma(f^*)$ for small t, which means that f^* is not an optimum.

The following proposition shows how the defect is connected with the "manifold question" discussed above.

PROPOSITION 6.4 (see [BRT] and [T]). *Suppose that the disc $f \in H^\infty_N \cap C^\alpha$ is attached to a smooth manifold $M = \{z \in \mathbf{C}^N : \rho_1(z) = \cdots = \rho_l(z) = 0\}$. Let E be the set of discs passing through point $f(0)$, i.e., $E = \{g \in H^\infty_N \cap C^\alpha : g(0) = f(0)\}$. If the defect of f is equal to zero, then the derivative map*

$$\mathcal{F}_0 : T_f E \to C^\alpha(\mathbf{T}, \mathbf{R}), \qquad \mathcal{F}_0(h) = 2\mathrm{Re}\left(\frac{\partial \rho(f)}{\partial z} h\right)$$

is onto. Here $T_f E$ denotes the tangent space to E at f, i.e., $T_f E = \{h \in H^\infty_N \cap C^\alpha : h(0) = 0\}$.

The proof of Proposition 6.4 is a line-by-line repetition of the proof of Lemma 3.1.

One can easily verify that Proposition 6.4 implies that the set of discs attached to M and passing through $f(0)$ forms a Banach manifold near f.

Acknowledgment. The authors would like to thank M. S. Baouendi for showing us Trepreau's preprint [T] and for several useful discussions.

REFERENCES

[BB] S. Boyd and C. Barratt, *Linear Controller Design*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[BG] E. Bedford and B. Gaveau, *Envelopes of holomorphy of certain two-spheres in C^2*, Amer. J. Math., 105 (1983), pp. 975–1009

[Bog] A. Boggess, *CR manifolds and the tangential Cauchy–Riemann complex*, CRC Press, Boca Raton, FL, 1991.

[BT] M. S. Baouendi and F. Treves, *A property of the functions and distributions annihilated by a locally integrable system of complex vector fields*, Ann. of Math., 113 (1981), pp. 341–421.

[BRT] M. S. Baouendi, L. P. Rothschild, and J.-M. Trepreau, *On the geometry of analytic discs attached to real manifolds*, J. Differential Geom., 39 (1994), pp. 379–405.

[D] J. Doyle, *The μ-Analysis and Synthesis Toolbox for Use with* MATLAB, documentation, The MathWorks, Inc., Natick, MA.

[FKTW] M. Fan, J. Koninckx, A. Tits, and L. Wang, *Tandem for optimization-based design interacting with arbitrary simulators*, report, Systems Research Center, University of Maryland, College Park, MD.

[F] F. Forstneric, *Analytic discs with boundaries in a maximal real submanifold of C^2*, Ann. Inst. Fourier (Grenoble), 37 (1987), pp. 1–44.

[Fr] B. Francis, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, Berlin, 1986.

[G] J. Globevnik, *Perturbation by analytic discs along maximal real submanifolds of $\mathbf{C}^N$*, Math. Z., to appear.

[H] J. W. Helton, *Operator theory, analytic functions, matrices, and electrical engineering*, in Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics 68, CBMS, Washington, DC, 1985.

[H1] J. W. Helton, *Optimization over spaces of analytic functions and the Corona problem*, J. Oper. Theory, 13 (1986), pp. 359–375.

[HH] J. W. Helton and R. Howe, *A bang-bang principle for the frequency domain*, J. Approx. Theory, 47 (1986), pp. 101–121.

[HMar] J. W. Helton and D. E. Marshall, *Frequency domain design and analytic selections*, Indiana Univ. Math. J., 39 (1990), pp. 157–184.

[HM1] J. W. Helton and O. Merino, *Conditions for optimality over H^∞,* SIAM J. Control Optim., 31 (1993), pp. 1379–1415.

[HMW] J. W. Helton, O. Merino, and T. Walker, *Algorithms for optimizing over analytic functions*, Indiana Univ. Math J., 42 (1993), pp. 839–874.

[MNPW] D. Q. Mayne, W. T. Nye, L. Polak, and T. Wu, *DELIGHT MIMO: An interactive, optimization based multivariable control system design package*, in Computer-Aided Control Systems Engineering, M. Jamshidi and C. J. Herget, eds., North–Holland, Amsterdam, 1985.

[O] Y.-G. Oh, *Fredholm-regularity and realization of Riemann–Hilbert problem and application to the perturbation theory of analytic discs*, preprint.

[PY] V. V. Peller and N. J. Young, *Superoptimal analytic approximations of matrix functions*, J. Funct. Anal., to appear.

[Si] T. Sideris, *Robust feedback synthesis via conformal mappings and H_∞ optimization*, Ph.D. thesis, University of Southern California, Los Angeles, 1985.

[St] R. L. Sreit, *Solution of systems of complex linear equations in the l_∞ norm with constraints on the unknowns,* SIAM J. Sci. Statist. Comput., 7 (1986), pp. 132–149.

[T] J.-M. Trepreau, *On the global Bishop equation,* preprint.

[Tu1] A. E. Tumanov, *Extension of CR functions into a wedge from a manifold of finite type,* Mat. Sb., 136(178) (1988), pp. 128–139 (in Russian); Math. USSR-Sb., 64 (1989), pp. 129–140 (in English).

[Tu2] A. E. Tumanov, *Extension of CR-functions into a wedge,* Mat. Sb., 181 (1990), pp. 951–964 (in Russian); Math. USSR-Sb., 70 (1991), pp. 385–398 (in English).

[Ve] N. P. Vekua, *Systems of Singular Integral Equations,* P. Noordhoff, Groningen, the Netherlands, 1967.

[V] A. E. Vityaev, *On the uniqueness of a H^∞ optimization problem,* J. Geom. Anal., to appear.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 769–807, July 1997

001

VOLUME-PRESERVING MEAN CURVATURE FLOW AS A LIMIT OF A NONLOCAL GINZBURG–LANDAU EQUATION[*]

LIA BRONSARD[†] and BARBARA STOTH[‡]

Abstract. We study the asymptotic behavior of radially symmetric solutions of the nonlocal equation

$$\varepsilon\varphi_t - \varepsilon\Delta\varphi + \frac{1}{\varepsilon}W'(\varphi) - \lambda_\varepsilon(t) = 0$$

in a bounded spherically symmetric domain $\Omega \subset \mathbf{R}^n$, where $\lambda_\varepsilon(t) = \frac{1}{\varepsilon}\fint_\Omega W'(\varphi)\,dx$, with a Neumann boundary condition. The analysis is based on "energy methods" combined with some a priori estimates, the latter being used to approximate the solution by the first two terms of an asymptotic expansion. We only need to assume that the initial data as well as their energy are bounded. We show that, in the limit as $\varepsilon \to 0$, the interfaces move by a nonlocal mean curvature flow, which preserves mass. As a by-product of our analysis, we obtain an L^2 estimate on the "Lagrange multiplier" $\lambda_\varepsilon(t)$, which holds in the nonradial case as well. In addition, we show rigorously (in general geometry) that the nonlocal Ginzburg–Landau equation and the Cahn–Hilliard equation occur as special degenerate limits of a viscous Cahn–Hilliard equation.

Key words. nonlocal mean curvature flow, nonlocal Allen–Cahn equation

AMS subject classifications. 35B25, 35K57

PII. S0036141094279279

1. Introduction.

We consider the nonlocal reaction-diffusion equation recently introduced by Rubinstein and Sternberg [RS],

$$\begin{aligned} &\varepsilon\varphi_t - \varepsilon\triangle\varphi + \frac{1}{\varepsilon}W'(\varphi) - \lambda_\varepsilon(t) = 0, \\ &\qquad\qquad\qquad \lambda_\varepsilon(t) = \frac{1}{\varepsilon}\fint_\Omega W'(\varphi)\,dx \end{aligned} \tag{1.1}$$

in a bounded domain $\Omega \subset \mathbf{R}^n$, $n \geq 2$, with Neumann boundary condition

$$\left.\frac{\partial\varphi}{\partial n}\right|_{\partial\Omega\times[0,T]} = 0. \tag{1.2}$$

The potential W is a bistable potential, that is, $W \geq 0$ and it vanishes exactly at two points. The typical bistable potential is given by

$$W(\varphi) = \frac{1}{2}(1-\varphi^2)^2, \tag{1.3}$$

and we will present our results for this specific potential. However, we point out that our results can be extended to the more general case.

[*] Received by the editors December 27, 1994; accepted for publication (in revised form) February 27, 1996. This research was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis and is available as Report 94-March-NA-008, Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA, 1994.

http://www.siam.org/journals/siam/28-4/27927.html

[†] Department of Mathematics, McMaster University, Hamilton, ON L8S 4K1, Canada (bronsard@math.mcmaster.ca). This author was supported by an NSERC (Canada) grant.

[‡] Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA 15213 and Institut für Angewandte Mathematik, Universität Bonn, 53115 Bonn, Germany (bstoth@iam.uni-bonn.de).

An important property of this flow is that its mass is preserved, that is,

$$\partial_t \int_\Omega \varphi(x,t)dx = 0. \tag{1.4}$$

Rubinstein and Sternberg [RS] introduced the nonlocal equation (1.1) as a simpler alternative to the classical Cahn–Hilliard equation [CH]

$$\varepsilon\varphi_t = \Delta\left(-\varepsilon\Delta\varphi + \frac{1}{\varepsilon}W'(\varphi)\right), \tag{1.5}$$

to model phase separation after quenching (rapid cooling) of homogeneous binary systems such as glasses and polymers. The function φ represents the difference in concentration of the two species of the binary mixture and hence is a conserved quantity. Using multiple-time-scale asymptotic expansions to study the behavior of the solution to (1.1)–(1.2), Rubinstein and Sternberg [RS] formally obtained that the domain Ω is divided in regions where φ is close to the local minima of W. Moreover, the interfaces $\{\Gamma_i\}$ dividing these regions evolve (in the limit $\varepsilon \to 0$) with normal velocity

$$V_i = \kappa_i - \sum \frac{1}{\sum |\Gamma_j|} \int_{\Gamma_j} \kappa_j,$$

where κ_i is the sum of principal curvatures of Γ_i and $|\Gamma_i|$ is its perimeter. This is a nonlocal volume-preserving mean curvature flow. We propose to use an energy-type method to rigorously justify this picture in a certain radially symmetric setting. More specifically, we assume that Ω is a ball in $\mathbf{R}^n$ and that φ is radial with several "transition" spheres. Equation (1.1) is already written in the time scale for which the above nonlocal mean curvature flow occurs in times of order 1. However, by rescaling, we see that this problem corresponds to the singular perturbation problem $\varphi_\tau - \varepsilon^2\Delta\varphi + W'(\varphi) - \varepsilon\lambda = 0$.

Next, we shall compare the two equations (1.1) and (1.5), as well as their respective asymptotic limiting flows. The Cahn–Hilliard equation is the gradient flow in the dual norm of some suitable subspace of $H^{1,2}(\Omega)$ (cf. [F]) for the functional

$$E_\varepsilon[\varphi] = \int_\Omega \frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{\varepsilon}W(\varphi)\,dx, \tag{1.6}$$

while the nonlocal equation (1.1) is the gradient flow in $L^2(\Omega)$ for the same functional (1.6) against the mass constraint (1.4). The associated (time-independent) minimization problem, that is, the problem of minimizing (1.6) with a mass constraint, has been studied by Luckhaus and Modica [LM]. They rigorously obtained the first-order expansion in ε of the associated Lagrange multiplier. In this context, we can loosely interpret the nonlocal term $\lambda = \lambda_\varepsilon(t)$ in (1.1) as a Lagrange multiplier. In fact, because of the Neumann boundary condition (1.2), the expression for $\lambda_\varepsilon(t)$ is exactly what is needed for the gradient flow of $E_\varepsilon[\varphi]$ to conserve mass.

The asymptotic behavior of the solutions to (1.1) and (1.5) are very different. The formal analysis of Pego [P] suggests that the asymptotic behavior of (1.5) is given by the so-called Mullins–Sekerka problem [MS] (sometimes called the Hele–Shaw problem):

$$\Delta u = 0 \quad \text{in } \Omega\backslash\Gamma, \qquad u = -\kappa \quad \text{on } \Gamma, \qquad \left[\frac{\partial u}{\partial n}\right]_\Gamma = -V,$$

where Γ is the interface, κ is the sum of its principal curvatures, and V is its normal velocity. This has recently been proved by Stoth [S2] in the radial case in $\mathbf{R}^n$, $n \leq 3$. Alikakos, Bates, and Chen [ABC] have a convergence result in general domains, assuming that the limit flow is smooth and with particular initial and boundary conditions.

Both limiting flows are nonlocal and some existence results are known for each of them. Indeed, Gage [Ga] (for curves) and Huisken [H] have proved that a convex manifold evolving by volume-constrained mean curvature flow eventually becomes a sphere with the prescribed area. Also, there are simple examples which show that nonconvex curves may develop singularities in finite time (see, e.g., [RS]). For the Mullins–Sekerka problem, Chen [C] has proved a weak, local-in-time existence result for general smooth initial curves and a global existence result for curves which are small perturbations of a circle.

The most striking difference between the two limiting geometric flows is the effect of small spheres. Indeed, in the radial case, we can easily calculate the respective evolution laws for the interfaces explicitly. In the three-dimensional case, assuming that there are two interfaces $r_2(t) < r_1(t)$, the nonlocal problem is given by (cf. (6.25))

$$\dot{r}_1 = 2\left(-\frac{1}{r_1} + \frac{r_1 - r_2}{r_1^2 + r_2^2}\right), \qquad \dot{r}_2 = 2\left(-\frac{1}{r_2} - \frac{r_1 - r_2}{r_1^2 + r_2^2}\right), \tag{1.7}$$

while the Mullins–Sekerka problem is given by

$$\dot{r}_1 r_1^2 = \dot{r}_2 r_2^2, \qquad \dot{r}_1 = \frac{-2}{r_1^2}\frac{r_1 + r_2}{r_1 - r_2}. \tag{1.8}$$

Therefore, as r_2 approaches 0, it is clear that $\dot{r}_1$ approaches 0 in (1.7), while it approaches $\frac{-2}{r_1^2}$ in (1.8). However, once the smallest sphere has disappeared, $\dot{r}_1$ must be 0 since mass must be preserved. This means that the flow for r_1 in the Mullins–Sekerka model is strongly affected by asymptotically small spheres. In fact, Rubinstein and Sternberg [RS] used a multiple scattering expansion known as the point-interaction approximation method to suggest that the Mullins–Sekerka problem is not the appropriate approximation of the Cahn–Hilliard equation when there are asymptotically small spheres.

There is an interesting connection between equations (1.1) and (1.5). Indeed, Rubinstein and Sternberg [RS] observed that equations (1.1) and (1.5) arise by formally taking different parameter limits ($\alpha \to 0$ and $\nu \to 0$, respectively) in the viscous Cahn–Hilliard equation $\alpha\varphi_t = \Delta(W'(\varphi) - \beta\Delta\varphi + \nu\varphi_t)$. This equation was introduced by Novick-Cohen [NC] in order to include viscous effects in the Cahn–Hilliard model. We prove these convergence results rigorously in section 7. This suggests that by taking an appropriate choice of the parameter limit in the viscous Cahn–Hilliard equation, one should recover a different limit flow with possibly better properties.

We note that the singular limit of (1.1) provides a notion of a weak solution for the nonlocal mean curvature flow. However, there is no uniqueness theorem in general: different sequences of ε's might produce different limits. The same approach has been used to define a model for mean curvature flow (cf. [BK2], [DS1], and [DS2]) using the Allen–Cahn equation [AC]. In that case, Evans, Soner, and Souganidis [ESS] have shown that this model coincides with the notion of motion by mean curvature flow in the sense of viscosity solutions (cf. [CGG] and [ES]).

We prove the convergence of the nonlocal equation (1.1) to volume-preserving mean curvature flow in a radially symmetric setting. We assume that for $\varphi = \varphi_\varepsilon$,

$$\begin{aligned} \|\varphi(\cdot, 0)\|_\infty &\leq C_0, \\ E_\varepsilon[\varphi](0) &\leq C_0, \\ \left| \int_\Omega \varphi(x, 0)\, dx \right| &< |\Omega| - \omega \end{aligned} \tag{1.9}$$

for some positive constants C_0 and ω. The second assumption means that the initial data must have a "transition-layer structure," i.e., $\varphi_\varepsilon \approx \pm 1$. The third condition ensures that there exists at least one interface. The case of general initial data is much harder; we refer to Soner [So] for the equivalent problem for the Allen–Cahn equation. Our method is an energy-type method similar to the methods developed by Bronsard and Kohn [BK1], [BK2] in order to study the singular limit of the Allen–Cahn equation and the methods developed by Stoth [S1], [S2] in order to study the singular limit of the phase-field model and the Cahn–Hilliard equation. The new feature here is the nonlocal nature of the equation, which does not allow for a comparison principle, and there does not seem to be a monotonicity formula.

We now describe the method in more detail. We first use BV bounds (Proposition 2.1) to obtain the existence of an L^1 limit v for a subsequence of φ_ε (Remark 2.4), and then we restrict our discussion to this subsequence. In addition, we show a uniform L^2 estimate on the Lagrange multiplier (Proposition 2.3), which implies the existence of a weak limit λ_0 for an appropriate subsequence. Next, we establish a monotone L^1 limit E_0 for $E_\varepsilon[\varphi_\varepsilon]$ (Corollary 2.5) which is used to define time intervals on which the variation of $E_\varepsilon[\varphi_\varepsilon]$ is uniformly small (Lemma 2.6). These results are not restricted to the radially symmetric case.

The next step is the foundation of our approach. We show that away from the origin and except at finitely many time points, φ_ε is close to the stationary-wave solution associated with the equation $\partial_t u - \triangle u + W'(u) = 0$. In the typical case where W is given by (1.3), the stationary-wave solution is tanh. More precisely, we obtain a locally uniform-in-time bound on $|| - \varepsilon^2 |\varphi_\varepsilon'|^2 + 2W(\varphi_\varepsilon)||_{L^\infty(R_0, 1)}$ which is valid except at finitely many time points (Proposition 3.2). Since $W(\varphi_\varepsilon)$ is bounded away from 0 in the transition region of φ_ε, this means that $|\varphi_\varepsilon'|$ is strictly bounded away from 0 in that region. Therefore, using the implicit function theorem, the level sets of φ_ε are given by Hölder-$\frac{1}{2}$ graphs $r = r_\varepsilon^i(t)$ (see (4.4)) that converge to some limits $r = \bar{r}^i(t)$ (see (4.5)). The task is to find the evolution equation satisfied by $r = \bar{r}^i(t)$.

We present the idea of the method for the ith interface $\bar{r} = \bar{r}^i$ and its approximation $r_\varepsilon = r_\varepsilon^i$. Let $z = \frac{r - r_\varepsilon}{\varepsilon}$ be a rescaling and $\Phi(z, t) = \varphi(r, t)$. The equation for Φ becomes

$$\varepsilon \partial_t \Phi - \dot{r}_\varepsilon \Phi' - \frac{1}{\varepsilon} \Phi'' - \frac{n-1}{\varepsilon z + r_\varepsilon} \Phi' + \frac{1}{\varepsilon} W'(\Phi) - \lambda_\varepsilon = 0.$$

We multiply it by $\Phi' \zeta$, where ζ is a smooth time-dependent test function, and in order to localize around r_ε, we integrate over $(-\frac{1}{\sqrt{\varepsilon}}, \frac{1}{\sqrt{\varepsilon}})$ and over (t_1, t_2). This gives

$$\begin{aligned} &\iint \zeta \left[\varepsilon \Phi_t \Phi' - \dot{r}_\varepsilon (\Phi')^2 - \frac{n-1}{\varepsilon z + r_\varepsilon} (\Phi')^2 - \lambda_\varepsilon \Phi' \right] dz\, dt \\ &\quad = \int \zeta \frac{1}{\varepsilon} \left[\frac{1}{2} (\Phi')^2 - W(\Phi) \right]_{-\frac{1}{\sqrt{\varepsilon}}}^{\frac{1}{\sqrt{\varepsilon}}} dt. \end{aligned}$$

Now if $\Phi(z)$ were the expected stationary-wave solution $\pm\tanh(z) =: \Phi_0(z)$, this would lead to the following equation for the limit $\bar{r}$:

$$-c_0 \int \zeta \left[\dot{\bar{r}} + \frac{(n-1)}{\bar{r}}\right] dt = 2 \int \zeta \nu(\bar{r}) \lambda_0 \; dt,$$

where $c_0 = \int_{-1}^{1} \sqrt{2W(\varphi)}\, d\varphi$ is the constant surface tension and ν is the direction of the jump. Thus all interfaces $\bar{r}^i$ evolve according to $-c_0(\dot{\bar{r}}^i + \frac{n-1}{\bar{r}^i}) = 2\nu^i \lambda_0$ (Proposition 6.3). However, using the mass-conservation property, we can calculate λ_0 explicitly in terms of $\bar{r}^i$ (Proposition 6.5), thereby deducing the equation for the limiting interface.

This formal derivation was done assuming that $\Phi = \Phi_0$ around each interface. Section 5 is devoted to establishing $H^{1,2}$ and $H^{1,\infty}$ bounds on the difference between Φ and Φ_0 (Corollary 5.9). We need a bound of order $\varepsilon^{\frac{1}{2}+s}$ for some positive s in order to replace Φ by Φ_0 in the above equation. As expected for this type of problem, this means that we have to prove these estimates for a higher-order expansion. It turns out that in our case a second-order expansion is sufficient (Proposition 5.5). The main observation used in the proof is that the linearization of the nonlinear operator around the stationary wave defines a strictly elliptic operator (see Berger and Fraenkel [BF] and Proposition 5.6).

We put everything together in section 6, where we rigorously derive the equation for the limiting interfaces. There are several difficulties to overcome. The most serious one is that we cannot exclude a priori the possibility that several ε-interfaces r_ε^i converge to the same limiting interface as $\varepsilon \to 0$, thereby giving rise to "multiplicities" higher than 1. There are two cases to distinguish. Either an odd number of ε-interfaces form a single limit interface, and this corresponds to a "true" interface, i.e., a jump of the limit v and $\nu = \pm 1$, or an even number of ε-interfaces form a single limit interface, and this corresponds to a "phantom" interface, i.e., $\nu = 0$. The latter correspond to interfaces separating the same phase. We prove that true interfaces evolve by the nonlocal flow and that their multiplicity is 1. Furthermore, we show that phantom interfaces evolve by mean curvature flow, but we do not characterize their multiplicity (Theorem 6.6). We point out that we do not establish the existence of phantom interfaces but only derive their law of motion. For the radial Allen–Cahn equation, it is possible to construct initial data that produce phantom interfaces in the limit (cf. [BS]), whereas for the Cahn–Hilliard equation, no phantom interfaces occur (cf. [S2]). Making use of the properties of nonlocal flow, we also show that all interfaces decrease and that at most two true interfaces can meet or nucleate at a given time point (Theorem 6.6). In fact, there are examples where two interfaces collide and disappear in the interior of the domain (Example 6.8). It is not clear whether they continue as a phantom interface or completely disappear. Their presence does not have an impact on the limit flow, but it accounts for an energy loss in the limit (Remark 6.9).

In the case where $n = 2$, we note that as long as there are an even number of interfaces, the nonlocal flow is simply mean curvature flow. In other words, the mean curvature flow preserves area in that case.

Our estimates of section 5 are strong enough to prove a formula for the limit energy E_0 that counts both true and phantom interfaces together with their multiplicities. From this it follows that there cannot be any nucleation in the interior if there is no nucleation at the origin. However, we cannot rule this out after the first geometric singularity of the nonlocal flow (see Remark 6.9).

Finally, we note that when $n = 1$, the evolution of the interfaces is expected to be exponentially slow in ε. This can easily be proven using the energy method of [BK1] combined with the result of Grant [G]. This exponentially slow motion has already been rigorously proven for the Cahn–Hilliard equation (cf. [ABF], [BH], [G], and [BX]).

2. Energy estimates. In this section, we derive all of the energy estimates necessary for the subsequent sections. We assume that φ_ε is a solution to (1.1) with the boundary data (1.2), that the domain $\Omega \subset \mathbf{R}^n$ is bounded with a Lipschitz boundary, and that for some $C_0 > 0$ and some $\omega > 0$ independent of ε,

$$E_\varepsilon[\varphi_\varepsilon(\cdot, 0)] \le C_0, \tag{2.1}$$

$$\sup_{x\in\overline{\Omega}} |\varphi_\varepsilon(x,0)| \le C_0 \quad \text{and} \quad |\Omega| - \left| \int_\Omega \varphi_\varepsilon(x,0)\, dx \right| \ge \omega. \tag{2.2}$$

Throughout this paper, C will denote positive constants, that depend only on the space dimension n, the size $|\Omega|$, and the final time T as well as on C_0 and ω. We show that the energy

$$E_\varepsilon[\varphi] = \int_\Omega \frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{\varepsilon}W(\varphi)\, dx$$

is a Lyapunov functional for (1.1)–(1.2), and we use this fact to obtain appropriate BV bounds as well as some “weak” Hölder estimates on φ_ε. Next, we derive an L^2 bound of the Lagrange multiplier. We then produce an L^1 limit of the solution φ_ε and a weak L^2 limit of the Lagrange multiplier λ_ε. In addition, we use the fact that $E_\varepsilon[\varphi]$ is a monotone function to show that it is weakly compact in $\mathrm{BV}(0,T)$ and compact in $L^1(0,T)$. Finally, we use the monotonicity of the energy to construct positive time intervals, where the variation of the energy is uniformly small in ε.

Proposition 2.1 (energy estimates). *Let $\varphi := \varphi_\varepsilon$ be a solution to* (1.1) *with boundary condition* (1.2) *and suppose that the initial data satisfy* (2.1). *Let g be defined via $g'(s) := \sqrt{2W(s)}$ with $g(0) := 0$, and let $0 \le s < \tau \le T$. Then the following statements hold:*

$$\varepsilon \int_s^\tau \int_\Omega |\partial_t\varphi|^2\, dx\, dt + E_\varepsilon[\varphi](\tau) - E_\varepsilon[\varphi](s) = 0, \tag{2.3}$$

$$\sup_{t\in[0,T]} \int_\Omega |\nabla g(\varphi)|\, dx \le \sup_{t\in[0,T]} E_\varepsilon[\varphi](t) \le C \quad \text{(energy bound)}, \tag{2.4}$$

$$\int_s^\tau \int_\Omega |\partial_t g(\varphi)|\, dx\, dt \le C\sqrt{\tau - s}. \tag{2.5}$$

Proof. First, multiplying equation (1.1) by $\partial_t\varphi$ and integrating in x, it follows that

$$\varepsilon \int_\Omega |\partial_t\varphi|^2\, dx - \varepsilon\int_\Omega \partial_t\varphi\Delta\varphi\, dx + \frac{1}{\varepsilon}\int_\Omega \partial_t W(\varphi)\, dx - \lambda_\varepsilon(t)\int_\Omega \partial_t\varphi\, dx = 0.$$

Using the mass-conservation property (1.5) and integrating by parts, this reduces to

$$\varepsilon\int_\Omega |\partial_t\varphi|^2\, dx + \int_\Omega \partial_t\left[\frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{\varepsilon}W(\varphi)\right]\, dx = 0.$$

Equation (2.3) follows from the definition of E_ε in (1.6) and integration in time from s to τ.

Next, we obtain the BV estimates (2.4) and (2.5). These estimates are not new (see, for example, [M] and [BK1] or [BK2]), but we include them for the sake of completeness. Using the definition of g, we have (cf. [M])

$$\int_\Omega |\nabla g(\varphi)|\,dx = \int_\Omega \sqrt{2W(\varphi)}|\nabla\varphi|\,dx \le E_\varepsilon[\varphi](t).$$

Inequality (2.4) now follows from (2.3) and the initial bound on the energy (2.1). Moreover, we obtain the "weak" Hölder estimate (cf. [BK2])

$$\begin{aligned}\int_s^\tau\int_\Omega |\partial_t g(\varphi)|\,dx\,dt &\le \left(\int_s^\tau\int_\Omega \frac{1}{\varepsilon}W(\varphi)\,dx\,dt\right)^{\frac12}\left(\int_s^\tau\int_\Omega \varepsilon|\partial_t\varphi|^2\,dx\,dt\right)^{\frac12}\\ &\le C\sqrt{\tau-s}.\end{aligned}$$

This completes the proof of Proposition 2.1. □

In particular, it follows that the functional E_ε is a Lyapunov functional for (1.1)–(1.2). From this fact follows the existence of a limit for φ_ε and an a priori bound on the Lagrange multiplier $\lambda_\varepsilon(t)\ \fint_\Omega W'(\varphi)\,dx$.

COROLLARY 2.2. *Under the same hypothesis as in Proposition* 2.1*, the following results hold:*

(2.6) $$\varphi \in L^\infty(0,T,L^4(\Omega)),$$

(2.7) $$\sup_t \lambda_\varepsilon(t) \le \frac{C}{\sqrt{\varepsilon}},$$

(2.8) $$\sup_{t,x}|\varphi(t,x)| \le C\sqrt{\varepsilon} + \sup_x|\varphi(x,0)| + 1.$$

Proof. Statement (2.6) is a direct consequence of Proposition 2.1. Inequality (2.7) follows from (2.6) since

$$\sup_t \lambda_\varepsilon = \sup_t \frac{1}{\varepsilon}\fint_\Omega 2(\varphi^3-\varphi)\,dx \le \sup_t \frac{C}{\varepsilon}\left(\int_\Omega \varphi^2\,dx\right)^{\frac12}\left(\int_\Omega W(\varphi)\,dx\right)^{\frac12} \le \frac{C}{\sqrt{\varepsilon}}.$$

Estimate (2.8) is a consequence of the maximum principle and (2.7). □

PROPOSITION 2.3 (estimate on the Lagrange multiplier). *Let* $\varphi = \varphi_\varepsilon$ *be as in Proposition* 2.1 *and assume in addition that* φ_ε *satisfies* (2.2)*. Then for* λ_ε *as in* (1.1),

$$\int_0^T |\lambda_\varepsilon(t)|^2\,dt \ \le\ C.$$

Proof. We multiply the differential equation (1.1) by $\nabla\varphi\cdot\zeta$, where ζ is a smooth function with values in $\mathbf{R}^n$. If $\zeta\cdot n = 0$ on $\partial\Omega$, we find that

$$\begin{aligned}\varepsilon\int_\Omega \partial_t\varphi\,\zeta\cdot\nabla\varphi\,dx + \varepsilon\int_\Omega \partial_j\varphi\partial_i\zeta_j\partial_i\varphi\,dx - \int_\Omega\left(\frac{\varepsilon}{2}|\nabla\varphi|^2+\frac{1}{\varepsilon}W(\varphi)\right)\mathrm{div}\zeta\,dx\\ = -\lambda_\varepsilon\int_\Omega \varphi\mathrm{div}\zeta\,dx.\end{aligned}$$

We extend φ by reflection to some neighborhood of Ω, and we choose a smooth Dirac sequence ψ_δ. We define $\xi = \xi_\varepsilon^\delta$ to be the solution of

$$\begin{cases} -\Delta\xi &= \varphi(t,\cdot) * \psi_\delta - \fint_\Omega \varphi(t,\cdot) * \psi_\delta \quad \text{in } \Omega, \\ -\dfrac{\partial \xi}{\partial n} &= 0 \quad \text{on } \partial\Omega, \\ \fint_\Omega \xi &= 0. \end{cases}$$

Elliptic regularity theory and (2.8) give

$$||\xi||_{C^{2,\alpha}(\overline{\Omega})} \le ||\varphi(t,\cdot) * \psi_\delta - \fint_\Omega \varphi(t,\cdot) * \psi_\delta||_{C^{0,\alpha}(\overline{\Omega})} \le C(\delta)||\varphi(\cdot,t)||_{L^\infty(\Omega)} \le C(\delta).$$

We now set $\zeta := \nabla\xi$. The energy bound (2.4) and the L^∞ bound (2.8) then imply

$$|\lambda_\varepsilon(t)| \left| \int_\Omega \varphi(\varphi * \psi_\delta - \fint_\Omega \varphi * \psi_\delta)\, dx \right| (t) \le C \left(\varepsilon \int_\Omega |\partial_t \varphi|^2\, dx \right)^{\frac{1}{2}} (t) + C(\delta).$$

However, using the conservation-of-mass property (1.4) and the energy bound (2.4), we obtain

$$\begin{aligned} &\left| \int_\Omega \varphi\Big(\varphi * \psi_\delta - \fint_\Omega \varphi * \psi_\delta\Big)\, dx \right| (t) \\ &\quad = |\Omega| + \int_\Omega (\varphi^2(x,t) - 1)\, dx - \int_\Omega (\varphi - \varphi * \psi_\delta)\varphi(x,t)\, dx \\ &\qquad - \frac{(\int_\Omega \varphi(x,t)\, dx)^2}{|\Omega|} + \frac{\int_\Omega \varphi(x,t)\, dx}{|\Omega|} \int_\Omega (\varphi - \varphi * \psi_\delta)(x,t)\, dx \\ &\quad \ge \frac{1}{|\Omega|}\left(|\Omega|^2 - \Big(\int_\Omega \varphi(x,0)\, dx \Big)^2 \right) - C\Big(\int_\Omega W(\varphi)(x,t)\, dx \Big)^{\frac{1}{2}} \\ &\qquad - C\int_\Omega |\varphi - \varphi * \psi_\delta|(x,t)\, dx \\ &\quad \ge 2\omega - C\sqrt{\varepsilon} - C \sup_{|h|\le\delta} \int_\Omega |\varphi(x+h,t) - \varphi(x,t)|\, dx. \end{aligned}$$

We now use the following result from [S2]: the energy bound (2.4) implies that $\sup_{|h|\le\delta} \int_\Omega |\varphi(x+h,t) - \varphi(x,t)|\, dx \le C\sqrt{\delta}$. Thus we eventually obtain

$$(2\omega - C\sqrt{\delta} - C\sqrt{\varepsilon})|\lambda_\varepsilon(t)| \le C \left(\varepsilon \int_\Omega |\partial_t \varphi|^2\, dx \right)^{\frac{1}{2}} (t) + C(\delta).$$

Taking the square of this expression, integration in time and (2.3) then yield the desired assertion. □

Remark 2.4. The energy estimates imply weak compactness for the sequence $g(\varphi_\varepsilon)$ in $BV(\Omega \times (0,T))$, so we can choose a subsequence $g(\varphi_\varepsilon) \xrightarrow[\varepsilon\to 0]{*} g(v)$ in BV. This in turn implies that for some subsequence $\varphi_\varepsilon \to v$ in $L^1(\Omega \times (0,T))$, since g^{-1} exists and $\varphi_\varepsilon \in L^\infty(0,T;L^4(\Omega))$. In addition, $v = \pm 1$ almost everywhere and it is "weakly" Hölder continuous

$$\int_\Omega |v(x,\tau) - v(x,s)|\, dx \le C\sqrt{\tau - s} \quad \text{for } T \ge \tau \ge s \ge 0.$$

This Hölder continuity of v in L^1 is due to Bronsard and Kohn [BK2]. Here it is a consequence of (2.5) since this estimate carries over to the limit by lower semicontinuity. In addition, we will assume that the initial data $\varphi_\varepsilon(\cdot, 0)$ converge in $L^1(\Omega)$ to $v(\cdot, 0)$.

Next, due to the bound on the Lagrange multiplier, we may select a further subsequence such that λ_ε converges to λ_0 weakly in $L^2(0, T)$.

Henceforth, we will consider only this subsequence, and we will still denote it by $\varphi_\varepsilon, \lambda_\varepsilon$. In what follows, we will select still other subsequences of this one, but this does not have an impact on v and λ_0.

Another important consequence of Proposition 2.1 is that $E_\varepsilon[\varphi](\cdot)$ is monotone decreasing in t and hence weakly compact in $BV(0, T)$.

COROLLARY 2.5. *Let φ_ε be as in Proposition* 2.1. *Then $E_\varepsilon[\varphi_\varepsilon](\cdot)$ is weakly compact in $BV(0,T)$. Therefore, for an appropriate subsequence of ε's, there exists a function $E_0(\cdot)$ such that*

$$\begin{aligned} &E_\varepsilon[\varphi_\varepsilon](\cdot) \longrightarrow E_0(\cdot) \quad \textit{in } L^1(0,T) \textit{ and almost everywhere,} \\ \partial_t &E_\varepsilon[\varphi_\varepsilon](\cdot) \xrightarrow{*} \partial_t E_0(\cdot), \end{aligned} \tag{2.9}$$

where the weak $*$ convergence is in $[C^0(0,T)]'$.

Proof. By (2.4), $E_\varepsilon[\varphi_\varepsilon]$ is clearly uniformly bounded in $L^1(0, T)$. Moreover, identity (2.3) implies that $E_\varepsilon[\varphi_\varepsilon]$ is monotone decreasing and thus uniformly bounded in $BV(0, T)$ by assumption (2.1) on the initial data. Thus we may select a subsequence of ε's as claimed. □

The results of Modica [M], Modica and Mortola [MM], and Sternberg [S] show that E_ε Γ-converges to a functional E_*, which is defined on BV functions by

$$E_*[v] = c_0 \int_\Omega |\nabla v| \, dx, \tag{2.10}$$

where $c_0 := \int_{-1}^{1} \sqrt{2W(\varphi)} \, d\varphi$. In particular, it is easy to see that for almost all t

$$\liminf_{\varepsilon \to 0} E_\varepsilon[\varphi_\varepsilon](t) \geq E_*[v](t). \tag{2.11}$$

From Corollary 2.5, we cannot conclude that E_0 is $E_*[v]$, but we know that for almost all t, we have $E_0(t) \geq E_*[v](t)$.

In Corollary 2.5, we have shown that for almost all t,

$$E_\varepsilon[\varphi_\varepsilon](t) \to E_0(t).$$

We define for any $\eta > 0$ a set $N(\eta) \subset [0, T]$ as the set of all jump points of E_0 with height at least η:

$$N(\eta) := \left\{ t \mid \operatorname*{ess\,inf}_{s<t} E_0(s) - \operatorname*{ess\,sup}_{s>t} E_0(s) \geq \eta \right\}. \tag{2.12}$$

Then for any $\eta > 0$, the set $N(\eta)$ is finite since E_0 is monotone decreasing in an L^1 sense:

$$E_0(t) \geq E_0(s) \quad \text{for almost every } s \geq t. \tag{2.13}$$

In fact, since $E_\varepsilon[\varphi_\varepsilon](t) \leq C$ by (2.4), it follows that

$$\# N(\eta) \leq \frac{C}{\eta}. \tag{2.14}$$

For $t_0 > 0$, we define $T_\varepsilon(\eta, t_0) > 0$ by

$$\varepsilon \int_{(t_0 - T_\varepsilon(\eta,t_0))^+}^{t_0+T_\varepsilon(\eta,t_0)} \int_\Omega (\partial_t \varphi_\varepsilon)^2 \, dx \, dt = \eta. \tag{2.15}$$

The following lemma is very important to our approach. It is based on the fact that E_0 is monotone decreasing. It basically says that given any $t_0 \notin N(\eta)$, we can find an open interval $(t_0 - T_0(\eta, t_0), t_0 + T_0(\eta, t_0))$ on which the variation of the energy $E_\varepsilon[\varphi_\varepsilon](\cdot)$ is uniformly small in ε.

LEMMA 2.6. *Let $\varphi = \varphi_\varepsilon$ be as in Proposition* 2.1. *Let $0 < t_0 \notin N(\eta)$, where $N(\eta)$ is given by* (2.12), *and let $T_\varepsilon(\eta, t_0)$ be as in* (2.15). *Then there exists $T_0(\eta, t_0) > 0$ such that*

$$T_\varepsilon(\eta, t_0) > T_0(\eta, t_0), \qquad \textit{for } \varepsilon \le \varepsilon_0(\eta, t_0).$$

In particular,

$$E_\varepsilon[\varphi_\varepsilon](t_0 - T_0)^+ - E_\varepsilon[\varphi_\varepsilon](t_0 + T_0) \le \eta.$$

Proof. Suppose to the contrary that $T_\varepsilon \to 0$ for some subsequence. Then using (2.3) and the monotonicity of $E_\varepsilon[\varphi_\varepsilon]$, we have for almost any $\tau > 0$ that

$$\begin{aligned} 0 < \eta &= \lim_{\varepsilon \to 0} \varepsilon \int_{t_0 - T_\varepsilon(\eta,t_0)}^{t_0+T_\varepsilon(\eta,t_0)} \int_\Omega (\partial_t \varphi)^2 \, dx \, dt \\ &= \lim_{\varepsilon \to 0} (E_\varepsilon[\varphi](t_0 - T_\varepsilon(\eta, t_0)) - E_\varepsilon[\varphi](t_0 + T_\varepsilon(\eta, t_0))) \\ &\le \lim_{\varepsilon \to 0} (E_\varepsilon[\varphi](t_0 - \tau) - E_\varepsilon[\varphi](t_0 + \tau)) \\ &= E_0(t_0 - \tau) - E_0(t_0 + \tau). \end{aligned}$$

Thus by the choice of t_0,

$$0 < \eta \le \operatorname{ess\,inf}_{s<t_0} E_0(s) - \operatorname{ess\,sup}_{s>t_0} E_0(s) < \eta. \qquad \square$$

3. A first approximation. The subsequent sections will be restricted to radially symmetric solutions; without loss of generality, we will assume that Ω is the unit disk in $\mathbf{R}^n$, $n \ge 2$.

In radial coordinates $r = |x|$, the evolution for $\varphi = \varphi_\varepsilon(r, t)$ becomes

$$\begin{aligned} \varepsilon \partial_t \varphi - \varepsilon \varphi_{rr} - \frac{\varepsilon(n-1)}{r} \varphi_r + \frac{1}{\varepsilon} W'(\varphi) - \lambda_\varepsilon(t) = 0, \\ \varphi(r, 0) = \varphi_\varepsilon^0(r), \end{aligned} \tag{3.1}$$

where as explained in the introduction, we choose $W(\varphi) = \frac{1}{2}(1 - \varphi^2)^2$. Moreover, since we consider the case of a Neumann boundary condition and since φ is smooth,

$$\varphi'(1, t) = 0 \quad \text{and} \quad \varphi'(0, t) = 0. \tag{3.2}$$

Thus mass is preserved. We assume the following conditions on the initial data. There exist some $C_0 > 0$ and some $\omega > 0$ such that for all ε,

$$\|\varphi_\varepsilon^0\|_{L^\infty(0,1)} \le C_0, \tag{3.3}$$

$$E_\varepsilon[\varphi_\varepsilon^0] \le C_0, \tag{3.4}$$

and

$$\left| \int_\Omega \varphi_\varepsilon^0 \, dx \right| < |\Omega| - \omega. \tag{3.5}$$

The following proposition is essential to the approach used in this paper. It is used to show that, away from the origin, the solution φ_ε is a priori close to the function $\pm q(\frac{\xi}{\varepsilon})$, where q solves

$$q_{\xi\xi} = W'(q) \quad \text{ with } \; q(-\infty) = -1, \quad q(\infty) = 1, \quad q(0) = 0.$$

In other words, the solution φ_ε is close to the one-dimensional stationary-wave solution $\pm q(\xi)$ associated with the equation $u_t = u_{\xi\xi} - W'(u)$, as is predicted by the formal asymptotic expansions of Rubinstein and Sternberg (cf. [RS]). For the existence and properties of the stationary-wave solution q, we refer to Aronson and Weinberger [AW] and Fife and McLeod [FM]. When $W(\varphi) = \frac{1}{2}(1-\varphi^2)^2$, this stationary wave is given by $q(\xi) = \tanh(\xi)$.

PROPOSITION 3.1. *Let $\varphi = \varphi_\varepsilon$ and φ_ε^0 satisfy* (3.1)–(3.5). *Let* $0 < R_0 < 1$ *and* $t_1 > t_2$. *Then for any* $t_2 < t < t_1$,

$$\left\| -\frac{\varepsilon^2}{2} |\varphi'|^2 + W(\varphi) \right\|_{L^\infty(R_0,1)} (t)$$
$$\le C(R_0) \left(\sqrt{\varepsilon} + \left(\varepsilon \int_{t_2}^{t_1} \int_\Omega \varphi_t^2 \, dx \, dt \right)^{\frac{1}{2}} + \left(\varepsilon^3 \int_\Omega \varphi_t^2(x, t_2) \, dx \right)^{\frac{1}{2}} \right).$$

Proof. First, we note that $||\varphi||_{L^\infty(\Omega\times(0,T))} \le C$ by assumption (3.3) and Corollary 2.2. Therefore, multiplying (3.1) by $\varepsilon\varphi'$, integrating over $(\eta, \rho) \subset (R_0, 1)$, using the fact that the energy is bounded (cf. (2.4)), and using the bound on $\lambda_\varepsilon(\cdot)$ (cf. (2.8)), it follows that

$$\begin{aligned} (3.6) \qquad & \left| -\frac{\varepsilon^2}{2} |\varphi'(\rho,t)|^2 + W(\varphi(\rho,t)) \right| \\ &= \Bigg| -\frac{\varepsilon^2}{2} |\varphi'(\eta,t)|^2 + W(\varphi(\eta,t)) - \varepsilon^2 \int_\eta^\rho \varphi_t \varphi' \, dr \\ &\quad + 2\varepsilon^2 \int_\eta^\rho \frac{1}{r} |\varphi'|^2 \, dr + \varepsilon\lambda_\varepsilon(t)(\varphi(\rho,t) - \varphi(\eta,t)) \Bigg| \\ &\le \frac{\varepsilon^2}{2} |\varphi'(\eta,t)|^2 + W(\varphi(\eta,t)) + 2\varepsilon|\lambda_\varepsilon(t)| \, ||\varphi||_{L^\infty(\Omega\times(0,T))} \\ &\quad + \frac{\varepsilon^2}{R_0^{n-1}} \left(\int_\eta^\rho (\varphi_t)^{n-1} r^{n-1} \, dr \right)^{\frac{1}{2}} \left(\int_\eta^\rho |\varphi'|^2 r^{n-1} \, dr \right)^{\frac{1}{2}} + \frac{2\varepsilon^2}{R_0^n} \int_\eta^\rho |\varphi'|^2 r^{n-1} \, dr \\ &\le \frac{\varepsilon^2}{2} |\varphi'(\eta,t)|^2 + W(\varphi(\eta,t)) + C \frac{\varepsilon^{\frac{3}{2}}}{R_0^{n-1}} \left(\int_{R_0}^1 (\varphi_t)^2 r^{n-1} \, dr \right)^{\frac{1}{2}} + C\frac{\varepsilon}{R_0^n} + C\sqrt{\varepsilon}. \end{aligned}$$

Next, integrating in η over $(R_0, 1)$ and again using the bound on the energy (2.4), we find for $t_0 \le t \le t_1$ that

$$\left| -\frac{\varepsilon^2}{2} |\varphi'(\rho,t)|^2 + W(\varphi(\rho,t)) \right| \le C(R_0) \left(\sqrt{\varepsilon} + \left(\varepsilon^3 \int_\Omega (\varphi_t(x,t))^2 \, dx \right)^{\frac{1}{2}} \right). \tag{3.7}$$

Thus we are left with estimating the last term in (3.7). For this we follow Stoth [S1] and consider the equation satisfied by $\partial_t\varphi$ on $\Omega \times (t_2, t_1)$:

$$\varepsilon\partial_{tt}\varphi - \varepsilon\Delta\partial_t\varphi + \frac{1}{\varepsilon}W''(\varphi)\partial_t\varphi - \partial_t\lambda_\varepsilon = 0.$$

We multiply this by $\varepsilon\partial_t\varphi$ and integrate over $\Omega\times(t_2,\tau)$ for $\tau < t_1$ to obtain

$$\begin{aligned}&\int_{t_2}^{\tau}\int_\Omega \varepsilon^2\varphi_{tt}\,\varphi_t\,dx\,dt - \int_{t_2}^{\tau}\int_\Omega \varepsilon^2\varphi_t\Delta\varphi_t\,dx\,dt\\ &\quad= -\int_{t_2}^{\tau}\int_\Omega W''(\varphi)(\varphi_t)^2\,dx\,dt + \int_{t_2}^{\tau}\varepsilon\partial_t\lambda_\varepsilon\int_\Omega\varphi_t\,dx\,dt = -\int_{t_2}^{\tau}\int_\Omega W''(\varphi)(\varphi_t)^2\,dx\,dt,\end{aligned}$$

by the mass-conservation property (1.4). Next, we integrate by parts and use the boundary condition (3.2) and the fact that $W''(\varphi) = 2(3\varphi^2-1)$ is bounded:

$$\begin{aligned}\int_\Omega \frac{\varepsilon^2}{2}|\varphi_t(x,\tau)|^2\,dx + &\int_{t_2}^{\tau}\int_\Omega \varepsilon^2|\nabla\varphi_t|^2\,dx\,dt\\ &\leq \int_\Omega\frac{\varepsilon^2}{2}|\varphi_t(x,t_2)|^2\,dx + C\int_{t_2}^{\tau}\int_\Omega(\varphi_t(x,t))^2\,dx\,dt.\end{aligned}$$

The proposition now follows from (3.7). □

Now according to Lemma 2.6, we can choose T_0 small enough such that

$$\varepsilon\int_{t_0-T_0}^{t_0+T_0}\int_\Omega(\varphi_t)^2\,dx\,dt$$

is as small as desired if $t_0 \notin N(\eta)$. This means that, away from the origin, the solution φ_ε is as close as we want to the stationary wave q in (t_0-T_0, t_0+T_0). This is the content of the following important consequence of Proposition 3.1 and Lemma 2.6.

Proposition 3.2 (first approximation). *Let $\varphi = \varphi_\varepsilon$ and φ^0_ε satisfy* (3.1)–(3.5). *Let $0 < R_0 < 1$ and $\delta > 0$. Then there exists $\eta = \eta(\delta, R_0)$ such that for any $0 \neq t_0 \notin N(\eta)$, there exists $T_0 = T_0(\delta, R_0, t_0) > 0$ and $\varepsilon_0 = \varepsilon_0(\delta, R_0, t_0) > 0$ such that*

$$\sup_{t_0-T_0\leq t\leq t_0+T_0}\left\|-\frac{\varepsilon^2}{2}|\varphi'|^2 + W(\varphi)\right\|_{L^\infty(R_0,1)} \leq \delta^2 \quad \textit{for } \varepsilon \leq \varepsilon_0.$$

We then rename $N(\eta(\delta,R_0))$ as $N(\delta,R_0)$.

Proof. Define η via $\sqrt{\eta} = \frac{\delta^2}{2C(R_0)}$, with $C(R_0)$ as in Proposition 3.1, and choose $T_0 = T_0(\delta, R_0, t_0)$ to be as in Lemma 2.6. Then we use Proposition 3.1 with $t_1 = t_0+T_0$ and the mean value over $t_2 \in (t_0 - T_0, t_0 - \frac{T_0}{2})$ to obtain for $t_0 - \frac{T_0}{2} \leq t \leq t_0 + T_0$,

$$\begin{aligned}&\left\|-\frac{\varepsilon^2}{2}|\varphi'(\cdot,t)|^2 + W(\varphi(\cdot,t))\right\|_{L^\infty(R_0,1)}\\ &\leq C(R_0)\left(\sqrt{\varepsilon} + \left(\varepsilon\int_{t_0-T_0}^{t_0+T_0}\int_\Omega\varphi_t^2\,dx\,dt\right)^{\frac12} + \left(\varepsilon^3\int_\Omega\varphi_t^2(x,t_2)\,dx\right)^{\frac12}\right)\\ &\leq C(R_0)\left(\sqrt{\varepsilon} + \left(\varepsilon\int_{t_0-T_0}^{t_0+T_0}\int_\Omega\varphi_t^2\,dx\,dt\right)^{\frac12} + \left(\varepsilon^3\frac{2}{T_0}\int_{t_0-T_0}^{t_0-\frac{T_0}{2}}\int_\Omega\varphi_t^2(x,t_2)\,dx\,dt_2\right)^{\frac12}\right)\\ &\leq C(R_0)\left(\sqrt{\varepsilon} + \frac{\delta^2}{2C(R_0)} + \varepsilon\sqrt{\frac{2}{T_0}}\right) \leq \delta^2\end{aligned}$$

for $\varepsilon \leq \varepsilon_0(\delta, R_0, t_0)$. We then rename $\frac{T_0}{2}$ as T_0. □

Remark 3.3. If $t_0 = 0$, then the same result as Proposition 3.2 holds with $(-T_0, T_0)$ substituted by $[0, T_0)$. A condition for this is that $\varepsilon^3 \int_\Omega \partial_t \varphi_\varepsilon^2(x,0)\,dx \to 0$, which by equation (1.1) is equivalent to the condition $\varepsilon^3 \int_\Omega (\triangle\varphi_\varepsilon - \frac{1}{\varepsilon^2} W'(\varphi_\varepsilon))^2(x,0)\,dx \to 0$. This proposition is crucial to our approach. It has two important consequences. The first is that we can define the interfaces of φ_ε by showing that the level sets of φ_ε are graphs. This is done in section 4. The second consequence is an even better approximation of φ_ε by the stationary-wave solution associated to (1.1), namely, we obtain a second-order approximation for φ_ε. This will be shown in section 5. Finally, this approximation will be used to take the weak limit of equation (3.1) to obtain the desired limiting equation in section 6.

We conclude this section with a final estimate, which gives control over the surface area that vanishes at the origin.

PROPOSITION 3.4. *We have the additional estimate*

$$\text{(3.8)} \quad \int_0^T \int_0^1 |g(\varphi_\varepsilon)'| r^{n-2}\,dr\,dt \leq \int_0^T \int_0^1 \left(\frac{\varepsilon}{2}|\varphi_\varepsilon'|^2 \;+\; \frac{1}{\varepsilon}W(\varphi_\varepsilon)\right) r^{n-2}\,dr\,dt \;\leq\; C.$$

Proof. We multiply the nonlocal equation (3.1) by $(-r^{n-1}\varphi_\varepsilon')$ and integrate over $(0,s)$. This yields

$$\begin{aligned}\varepsilon \int_0^s \varphi_\varepsilon''\varphi_\varepsilon' r^{n-1} dr \;+\; (n-1)\varepsilon \int_0^s {\varphi_\varepsilon'}^2 r^{n-2} dr \;-\; \frac{1}{\varepsilon}\int_0^s W'(\varphi_\varepsilon)\varphi_\varepsilon' r^{n-1} dr \\ =\; -\lambda_\varepsilon \int_0^s \varphi_\varepsilon' r^{n-1} dr \;+\; \varepsilon \int_0^s \partial_t\varphi_\varepsilon \varphi_\varepsilon' r^{n-1} dr.\end{aligned}$$

Hence

$$\begin{aligned}&\frac{n-1}{2}\varepsilon \int_0^s {\varphi_\varepsilon'}^2 r^{n-2} dr \;+\; \frac{n-1}{\varepsilon}\int_0^s W(\varphi_\varepsilon) r^{n-2} dr \\ &\quad\leq\; \left(\frac{\varepsilon}{2}|\varphi_\varepsilon'|^2(s) \;+\; \frac{1}{\varepsilon}W(\varphi_\varepsilon)(s) \;+\; 2|\lambda_\varepsilon|\; ||\varphi_\varepsilon||_{L^\infty}\right) s^{n-1} \\ &\qquad+\; \varepsilon \int_0^s |\partial_t\varphi_\varepsilon \varphi_\varepsilon'|\; r^{n-1} dr.\end{aligned}$$

Now the left-hand side at $s = \frac{1}{2}$ is bounded by the mean value of the right-hand side taken over $s \in (\frac{1}{2}, 1)$. Therefore, the energy estimate (2.4), the L^∞ bound (2.6), and the bound on the Lagrange multiplier (2.3) give

$$\int_0^T \int_0^{1/2} \left(\frac{\varepsilon}{2}|\varphi_\varepsilon'|^2 \;+\; \frac{1}{\varepsilon}W(\varphi_\varepsilon)\right) r^{n-2}\,dr\,dt \;\leq\; C.$$

This establishes the result since in the interval $(\frac{1}{2}, 1)$, there is nothing to prove. □

4. Definition of the interfaces of φ_ε and of the limit v.

In this section, we present the definition and properties of the interfaces of φ_ε and of v. The definition of the interfaces is based on the fact that Proposition 3.2 implies a lower bound on $|\varphi_\varepsilon'|$ such that we can apply the implicit function theorem. Indeed, let

$$\text{(4.1)} \quad \delta^2 < \frac{1}{8} \quad \text{and} \quad 0 < Q < \tanh\left[\frac{1}{2} - \tanh^{-1}\frac{1}{\sqrt{3}}\right] \quad \text{be such that} \quad W(Q) \geq \frac{1}{4}.$$

We will study the level sets of φ_ε of value less than Q. This precise choice of Q is important in the ellipticity proposition (Proposition 6.4) and in the following.

According to Proposition 3.2, $\frac{\varepsilon^2}{2}|\varphi_\varepsilon'|^2 \geq \frac{1}{8}$ in the subset of $(R_0, 1)\times(t_0-T_0, t_0+T_0)$ defined by $|\varphi_\varepsilon| \leq Q$ since in this set $W(\varphi_\varepsilon) \geq \frac{1}{4}$. This means that φ_ε must be monotone in r on each connected component of this set.

We assume that $\varphi = \varphi_\varepsilon$ satisfies (3.1)–(3.5), and for any $R_0 > 0$, we define

$$A_{R_0} := \bigcup_{t_0 \notin N(\delta, R_0)} (t_0 - T_0(\delta, R_0, t_0), t_0 + T_0(\delta, R_0, t_0)), \tag{4.2}$$

where δ is a fixed constant to be chosen later. We remark that by definition A_{R_0} is open and that its complement has at most finitely many points, all of them in $N(\delta, R_0)$. We choose an increasing sequence of open sets $D = D_m$ with $\bigcup D_m = A_{R_0}$.

The set $\overline{D}$ and hence D can be covered by finitely many of the $(t_0-T_0(\delta, R_0, t_0), t_0+T_0(\delta, R_0, t_0))$'s that were used to define A_{R_0}. Thus on D, Proposition 3.2 implies that

$$\sup_D \| -\varepsilon^2|\varphi'|^2 + 2W(\varphi)\|_{L^\infty(R_0,1)} \leq 2\delta^2$$

for all $\varepsilon \leq \varepsilon_0(\delta, R_0, D)$.

We now consider the "ε problem" in the strip $(R_0, 1) \times D$.

Let Q be as in (4.1). Then on $\{|\varphi_\varepsilon| \leq Q\} \cap (R_0, 1) \times D$, we have $\varepsilon^2|\varphi_\varepsilon'|^2 \geq \frac{1}{4}$. Thus for any $-Q < a < Q$, the set $\{\varphi_\varepsilon(r, t) = a\} \cap (R_0, 1) \times D$ consists of a collection of graphs $r_\varepsilon^i(\cdot, a)$.

Moreover, by the implicit function theorem, the following identities hold:

$$\begin{aligned} \partial_t\varphi_\varepsilon(r_\varepsilon^i(t,a),t) + \partial_t r_\varepsilon^i(t,a)\partial_r\varphi_\varepsilon(r_\varepsilon^i(t,a),t) &= 0, \\ \partial_a r_\varepsilon^i(t,a)\partial_r\varphi_\varepsilon(r_\varepsilon^i(t,a),t) &= 1. \end{aligned}$$

Using the coarea formula, this implies an $H^{1,2}$ estimate on r_ε^i. Indeed,

$$\begin{aligned} \int_{\{|\varphi_\varepsilon(\cdot,t)|<Q, r>R_0\}} \frac{|\partial_t\varphi_\varepsilon(r,t)|^2}{|\varphi_\varepsilon'(r,t)|}\,dr &= \int_{-Q}^{Q}\int_{\{\varphi_\varepsilon(\cdot,t)=a, r>R_0\}} \frac{|\partial_t\varphi_\varepsilon(r,t)|^2}{|\varphi_\varepsilon'(r,t)|^2}\,d\mathcal{H}^0\,da \\ &= \int_{-Q}^{Q}\sum_i |\partial_t r_\varepsilon^i(t,a)|^2\,da. \end{aligned}$$

Thus

$$\begin{aligned} \int_D\int_{-Q}^{Q}\sum_i |\partial_t r_\varepsilon^i(t,a)|^2\,da\,dt &\leq \frac{\varepsilon}{\sqrt{\delta}}\int_D\int_{R_0}^{1}|\partial_t\varphi_\varepsilon(r,t)|^2\,dr\,dt \\ &\leq C(R_0,\delta). \end{aligned}$$

Thus we may choose $a_\varepsilon \in (-Q, Q)$ such that (for some bigger $C(R_0, \delta)$)

$$\int_D \sum_i |\partial_t r_\varepsilon^i(t, a_\varepsilon)|^2\,dt \leq C(R_0, \delta).$$

This in turn implies that the graphs $r_\varepsilon^i(\cdot, a_\varepsilon)$ are Hölder-$\frac{1}{2}$ by embedding.

We define the interfaces of the ε problem by

$$r_\varepsilon^i(t) := r_\varepsilon^i(t, a_\varepsilon). \tag{4.3}$$

We note that none of these interfaces hits the fixed boundary $\partial\Omega$ because on the fixed boundary $\varphi_\varepsilon' = 0$ and thus by Proposition 3.2, the values of φ_ε have to be close to ± 1. However, on the interfaces, the values of φ_ε are given by a_ε and hence are uniformly away from ± 1. Thus all of the interfaces exist as long as they do not hit $r = R_0$. This allows us to introduce the ordering

$$(4.4) \qquad r_\varepsilon^i : I_\varepsilon^i \subset D \longrightarrow (R_0, 1) \quad \text{ for } i = 1, \ldots, M_\varepsilon$$

with $r_\varepsilon^i > r_\varepsilon^{i+1}$ and $r_\varepsilon^i = R_0$ on $\partial I_\varepsilon^i \cap D$. In addition $\operatorname{sign}(\varphi_\varepsilon(t,1) - a_\varepsilon)$ is fixed in D.

PROPOSITION 4.1. *The number M_ε of graphs $r_\varepsilon^i(t)$ is finite and bounded independently of D.*

Proof. By definition, $\varphi_\varepsilon(r_\varepsilon^i(t), t) = a_\varepsilon$ and therefore there exist points $c_\varepsilon^i(t)$ such that

$$r_\varepsilon^i(t) < c_\varepsilon^i(t) < r_\varepsilon^{i+1}(t)$$

with the property that $\varphi_\varepsilon'(c_\varepsilon^i(t), t) = 0$. Now the estimates given by Proposition 3.2 imply

$$W(\varphi_\varepsilon(c_\varepsilon^i(t), t)) \le \delta^2,$$

and $\varphi_\varepsilon(c_\varepsilon^i(t), t)$ have opposite signs for consecutive i's. Consequently,

$$|g(\varphi_\varepsilon(c_\varepsilon^{i+1}(t), t)) - g(\varphi_\varepsilon(c_\varepsilon^i(t), t))| = |g(\varphi_\varepsilon(c_\varepsilon^{i+1}(t), t))| + |g(\varphi_\varepsilon(c_\varepsilon^i(t), t))| \ge C(\delta) > 0.$$

Thus

$$\begin{aligned} C(\delta) M_\varepsilon &\le \sum_{i=1}^{M_\varepsilon} |g(\varphi_\varepsilon(c_\varepsilon^{i+1}(t), t)) - g(\varphi_\varepsilon(c_\varepsilon^i(t), t))| \\ &\le \sum_{i=1}^{M_\varepsilon} \left| \int_{c_\varepsilon^i(t)}^{c_\varepsilon^{i+1}(t)} g(\varphi_\varepsilon)' \, dr \right| \le C(R_0), \end{aligned}$$

by the energy estimate (2.4). Thus M_ε is uniformly bounded. □

As a result of this proposition, for a subsequence of ε's (depending on R_0), the number $M_\varepsilon =: M_0$ must be constant, and for $i = 1, \ldots, M_0$, there exist

$$(4.5) \qquad \bar{r}^i : I_{R_0}^i \subset A_{R_0} \to (R_0, 1] \quad \text{such that} \quad r_\varepsilon^i \longrightarrow \bar{r}^i$$

weakly in $H_{\text{loc}}^{1,2}(I_{R_0}^i)$ and locally uniformly.

In view of this, we define for any $R_0 > 0$ the set of limit interfaces

$$(4.6) \qquad \Gamma_{R_0} := \{(\bar{t}_0, \bar{r}_0) \mid \bar{t}_0 \in A_{R_0} \text{ and } \bar{r}_0 = \bar{r}^i(\bar{t}_0) \text{ for some } i = 1, \ldots, M_0\}.$$

This set Γ_{R_0} contains the free boundary $\partial\{v = -1\}$, but it may contain more.

Next, we study the "ε problem" locally around any $(\bar{t}_0, \bar{r}_0) \in \Gamma_{R_0}$.

Let m_0 be given by the property that at time $\bar{t}_0$, there are exactly m_0 graphs r_ε^i which converge to $\bar{r}_0$, i.e.,

$$(4.7) \qquad r_\varepsilon^i(\bar{t}_0) \to \bar{r}_0, \quad k \le i \le m_0 + k - 1.$$

Since all the r_ε^i are uniformly Hölder-$\frac{1}{2}$, there exists a box

$$B := (\bar{t}_1, \bar{t}_2) \times (a, b)$$

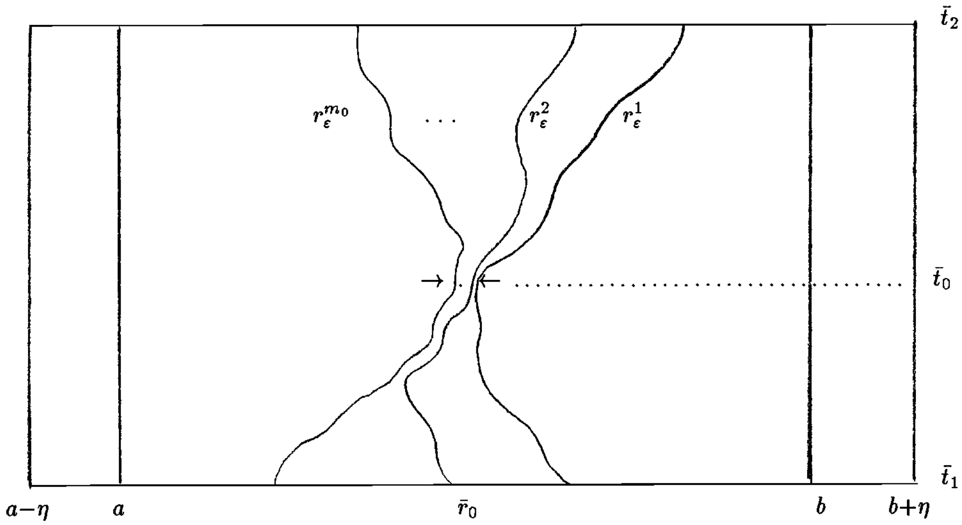


FIG. 1.

around $(\bar{t}_0, \bar{r}_0)$ that contains exactly m_0 graphs, all of them defined over the entire interval $(\bar{t}_1, \bar{t}_2)$. In addition, these graphs in B are $O(1)$ away from all other graphs. (See Figure 1.)

In conclusion, by putting together all of the above results, we have the following.

PROPOSITION/DEFINITION 4.2 (the local geometry). *For all $(\bar{t}_0, \bar{r}_0) \in \Gamma_{R_0}$ with $\bar{r}_0 \neq 1$, there exist natural numbers m_0 and k and a number $\nu \in \{-1, 0, +1\}$, a box $B = (\bar{t}_1, \bar{t}_2) \times (a, b) \subset D \times (R_0, 1)$, and $\eta > 0$ such that for $\varepsilon \leq \varepsilon_0(\bar{t}_0, \bar{r}_0)$, the following hold:*

(1) *$\{\varphi = a_\varepsilon\} \cap B$ consists of m_0 graphs r_ε^i over $(\bar{t}_1, \bar{t}_2)$, which are uniformly Hölder-$\frac{1}{2}$ and with derivatives uniformly in L^2, and $r_\varepsilon^i > r_\varepsilon^{i+1}$.*

(2) *At $\bar{t}_0$, exactly m_0 interfaces converge to $\bar{r}_0$.*

(3) *$r_\varepsilon^i \to \bar{r}^i$ uniformly and $\partial_t r_\varepsilon^i \rightharpoonup \partial_t \bar{r}^i$ weakly in $L^2(\bar{t}_1, \bar{t}_2)$ for $k \leq i \leq m_0+k-1$.*

(4) *$\bar{r}^i(\bar{t}_0) = \bar{r}_0$ for $k \leq i \leq m_0 + k - 1$.*

(5) *$a - \eta > R_0$ and the sets $\{(\bar{t}_1, \bar{t}_2) \times (a-\eta, a]\} \cap \{\varphi^\varepsilon = a_\varepsilon\}$ and $\{(\bar{t}_1, \bar{t}_2) \times [b, b+\eta)\} \cap \{\varphi^\varepsilon = a_\varepsilon\}$ are empty.*

(6)

$$\nu = \begin{cases} +1 & \textit{if } \varphi_\varepsilon(\bar{t}_0, a) < a_\varepsilon \textit{ and } \varphi_\varepsilon(\bar{t}_0, b) > a_\varepsilon, \\ -1 & \textit{if } \varphi_\varepsilon(\bar{t}_0, a) > a_\varepsilon \textit{ and } \varphi_\varepsilon(\bar{t}_0, b) < a_\varepsilon, \\ 0 & \textit{otherwise.} \end{cases}$$

If $\bar{r}_0 = 1$, then $B = (\bar{t}_1, \bar{t}_2) \times (a, b) \not\subset D \times (R_0, 1)$, but if we continue φ_ε by its boundary values, the above definitions remain meaningful.

Due to the symmetry of the argument, later we will explicitly describe only the case where $\varphi_\varepsilon(\bar{t}_0, b) > a_\varepsilon$ such that ν is either $+1$ or 0. The case where $\nu = 0$ corresponds to the case where the limit v has a "phantom" interface at which v "jumps" from 1 to 1 or -1 to -1, whereas the case where $\nu \neq 0$ corresponds to true interfaces of v.

5. A rigorous first-order expansion. Once again, throughout this section, we assume that φ_ε satisfies (3.1)–(3.5) such that the analysis of the preceding sections is valid.

We now have well-defined interfaces. We propose to study the solution near each interface. Our final goal is to pass to the limit in equation (3.1) around each interface. For this we will need a very good approximation of the solution φ_ε in $H^{1,\infty}$. This section is devoted to obtaining this approximation. The idea is to show that the asymptotic expansion is rigorous up to second order, at least in a weak sense. We will show this using appropriate $H^{1,2}$ error estimates. However, we will not prove an approximation of φ_ε everywhere in Ω as in [S1]. Instead, with the use of a cutoff function, we only consider the approximation of φ_ε locally around the interfaces.

In this section, we restrict our discussion to the box B defined in Proposition 4.2. We assume for simplicity of presentation that $k = 1$. First, we introduce a stretched variable around the largest interface $r^1_\varepsilon(t)$ in the box B. Let

$$z := \frac{|x| - r^1_\varepsilon(t)}{\varepsilon} \tag{5.1}$$

such that $z \in (\frac{-r^1_\varepsilon(t)}{\varepsilon}, \frac{1-r^1_\varepsilon(t)}{\varepsilon})$. Henceforth, we shall use *capital letters* for functions defined in the stretched variables and *lower-case letters* for functions written in the original variables so that, for example,

$$\Phi_\varepsilon(z,t) := \varphi_\varepsilon(r,t). \tag{5.2}$$

Moreover, the index ε will be dropped whenever it does not affect the clarity of the text. Then for η as in Proposition 4.2, the rescaling (5.1) maps the collection of points

$$a - \eta < a < r^{m_0}_\varepsilon < \cdots < r^2_\varepsilon < r^1_\varepsilon < b < b + \eta \tag{5.3}$$

onto

$$z^-_\varepsilon - \frac{\eta}{\varepsilon} < z^-_\varepsilon < z^{m_0}_\varepsilon < \cdots < z^2_\varepsilon < z^1_\varepsilon \ (= 0) < z^+_\varepsilon < z^+_\varepsilon + \frac{\eta}{\varepsilon}. \tag{5.4}$$

(See Figure 2.)

Now motivated by the formal analysis of [RS], we make the ansatz that Φ_ε is well approximated near $z = 0$ by

$$\Theta^\varepsilon(z,t) := \Phi^\varepsilon_0(z,t) + \varepsilon\Phi^\varepsilon_1(z,t) \quad \text{for } z \in \left(z^-_\varepsilon - \frac{\eta}{\varepsilon}, z^+_\varepsilon + \frac{\eta}{\varepsilon}\right). \tag{5.5}$$

The zeroth-order term $\Phi^\varepsilon_0(z,t)$ is given by

$$\Phi^\varepsilon_0(z,t) := \sum_{i=1}^{m_0} \Xi^\varepsilon_i(z,t) \underbrace{\tanh((-1)^{i+1}[z - z^i_\varepsilon(t)] + \mu_\varepsilon)}_{=:\Phi^\varepsilon_{0i}(z,t)}, \tag{5.6}$$

where $\mu_\varepsilon = \tanh^{-1} a_\varepsilon$ and Ξ_i is a partition of unity. More precisely, for $2 \le i \le m_0 - 1$, the function Ξ_i has support in $(\frac{z^{i+1}+z^i}{2} - 1, \frac{z^i+z^{i-1}}{2} + 1)$, while Ξ_1 has support in $(\frac{z^2}{2} - 1, \infty)$ and Ξ_{m_0} has support in $(-\infty, \frac{z^{m_0}+z^{m_0-1}}{2} + 1)$. Moreover, Ξ'_i has support in two disjoint intervals, each of length 2, given typically by $(\frac{z^i+z^{i+1}}{2} - 1, \frac{z^i+z^{i+1}}{2} + 1)$. (See Figure 2.)

Remark 5.1. We later prove that $|z_i - z_{i-1}|$ is uniformly larger than 2 (cf. Lemma 5.8) as a consequence of the first approximation proposition (Proposition 3.2), so the above partition of unity is meaningful.

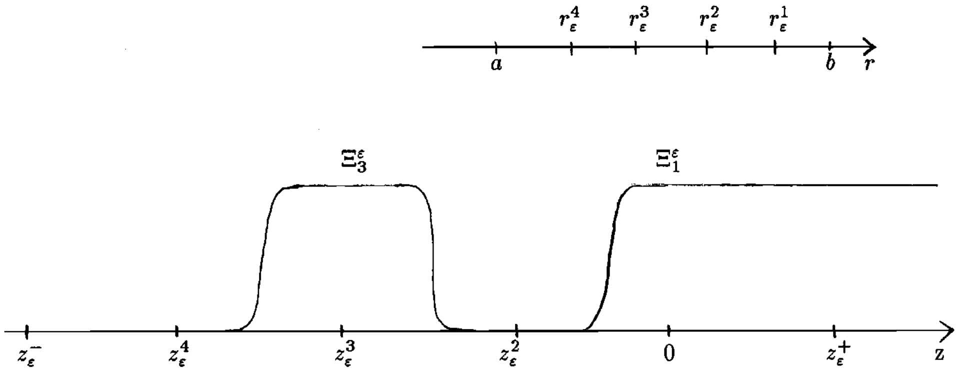


Fig. 2.

The first-order term is given by

$$\Phi_1^\varepsilon(z,t) := \sum_{i=0}^{m_0} \Xi_i^\varepsilon(z,t)\Phi_{1i}^\varepsilon(z,t), \tag{5.7}$$

where $\Phi_{1i} = \Phi_{1i}^\varepsilon(z,t)$ solves

$$\begin{cases} -\Phi_{1i}'' + W''(\Phi_{0i})\Phi_{1i} - \lambda_\varepsilon(t) = 0 & \text{in } (-\infty, z_\varepsilon^i) \cup (z_\varepsilon^i, \infty), \\ \Phi_{1i}\,(z_\varepsilon^i) = 0, \end{cases} \tag{5.8}$$

and $W''(\Phi) = 2(3\Phi^2 - 1)$. We note that equation (5.8) is the equation satisfied by the first-order term in the asymptotic expansion of [RS].

Remark 5.2. We do not impose that the differential equation for Φ_{1i} be satisfied at $z = z_\varepsilon^i$ in order to ensure that the solution remains uniformly bounded over the entire real axis (cf. Lemma 5.7). We refer to the work of Niethammer [N], who determined the expansion of the Lagrange multiplier for the radial, stationary problem with mass constraint by the condition that the equation be satisfied in the whole of $\mathbf{R}$.

Remark 5.3. The approximation depends on the direction of the jump. Here we give the definition for the case as selected in Proposition 4.2. If to the contrary the jump direction were the opposite, the tanh would have to be substituted by −tanh.

Remark 5.4. In the formal inner expansion, one also expands $\lambda_\varepsilon(t) = \lambda_0(t) + \varepsilon\lambda_1(t) + \cdots$ (see [RS]). Here we do not do this since we are interested only in the zeroth-order term. The lowest-order term will be determined later by the mass-conservation property.

The rest of this section is devoted to proving that (5.5) is indeed a good approximation to Φ_ε. To this end, let

$$\Psi^\varepsilon(z,t) := \Phi_\varepsilon(z,t) - \Theta^\varepsilon(z,t) \quad \text{for } \bar{t}_1 \le t \le \bar{t}_2. \tag{5.9}$$

Proposition 5.5 (first-order approximation). *Let ξ_ε be a cutoff function with*

$$\xi_\varepsilon(z,t) = \begin{cases} 1 & \textit{in } (-\varepsilon^{-\alpha} + z_\varepsilon^{m_0}(t), \varepsilon^{-\alpha}), \\ 0 & \textit{in } \mathbf{R}\backslash(-\varepsilon^{-\beta} + z_\varepsilon^{m_0}(t), \varepsilon^{-\beta}) \end{cases} \tag{5.10}$$

for $\frac{1}{2} < \alpha < \beta < 1$ so that

$$\text{supp}\ \xi(\cdot,t) \subset \left(z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon}\right).$$

Then for Ψ *given by* (5.9)*, we have the following estimates:*

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{(-\infty,z_\varepsilon^{m_0})\cup(0,\infty)}(|\Psi'|^2+|\Psi|^2)\xi^2\,dz\,dt\le C\varepsilon^{2\beta}, \tag{5.11}$$

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{z_\varepsilon^{m_0}}^{0}(|\Psi'|^2+|\Psi|^2)\,dz\,dt\to 0, \tag{5.12}$$

where C *depends on the data and the local box* B*.*

The rest of this section is devoted to the proof of this proposition. We first find the equation satisfied by Ψ^ε. Using the definition (5.1) for Φ_ε and equation (3.1), we find (if $z\le\frac{1-r_\varepsilon^1}{\varepsilon}$)

$$\begin{aligned}-\Phi''+W'(\Phi)-\varepsilon\lambda_\varepsilon(t)&=-\varepsilon^2\partial_t\varphi(\varepsilon z+r_\varepsilon^i(t),t)+\frac{2\varepsilon}{\varepsilon z+r_\varepsilon^1(t)}\Phi'\\&=:F_\varepsilon(t,z).\end{aligned}\tag{5.13}$$

We define $F_\varepsilon(t,z)$ by $W'(\varphi_\varepsilon(t,1))-\varepsilon\lambda_\varepsilon(t)$ for $z>\frac{1-r_\varepsilon^1}{\varepsilon}$.

The equation for Θ^ε is more complicated because of the extra terms that come from the partition of unity. To simplify the presentation, we let $\Theta_i:=\Phi_{0i}+\varepsilon\Phi_{1i}$. Then we have

$$-\Theta''+W'(\Theta)-\varepsilon\lambda_\varepsilon(t)=H_\varepsilon(z,t), \tag{5.14}$$

where

(5.15)

$$\begin{aligned}&H_\varepsilon(z,t):=\varepsilon^2\sum_{i=1}^{m_0}\Xi_i\left(6\Phi_{0i}\Phi_{1i}^2+2\varepsilon\Phi_{1i}^3\right)\\&+\sum_{i=1}^{m_0-1}[\Theta_{i+1}-\Theta_i]\left(\Xi_i''+2\Xi_i\Xi_{i+1}\{(1+\Xi_i)\Theta_i^2+(\Xi_i-2)\Theta_{i+1}^2+(1-2\Xi_i)\Theta_i\Theta_{i+1}\}\right)\\&+2\sum_{i=1}^{m_0-1}[\Theta_{i+1}-\Theta_i]'\,\Xi_i'.\end{aligned}$$

This formula comes from a linearization of $W'(\Phi)=-2\Phi(1-\Phi^2)$ around Φ_0. We note that this sum is only taken over two integers at a time because of the definition of Ξ_i. Also, as we will see later, the last two sums are of small order (cf. (5.25) and (5.26)).

Therefore, combining (5.13) and (5.14), the equation for the difference is

$$-\Psi''+W''(\Theta)\Psi=-2(3\Theta\Psi^2+\Psi^3)+F_\varepsilon(z,t)-H_\varepsilon(z,t), \tag{5.16}$$

and it holds in

$$\left(z_\varepsilon^- -\frac{\eta}{\varepsilon},z_\varepsilon^+ +\frac{\eta}{\varepsilon}\right)\backslash\{z_\varepsilon^1,\ldots,z_\varepsilon^{m_0}\} \tag{5.17}$$

with boundary values

$$\Psi^\varepsilon(z_\varepsilon^i,t)=0\quad\text{ for } i=1,\ldots,m_0. \tag{5.18}$$

Next, as in [S1], we follow an idea of Berger and Fraenkel and show that in some sense equation (5.16) for Ψ is uniformly strictly elliptic.

PROPOSITION 5.6 (ellipticity). *There exist $\zeta_1 > 0$ and $\zeta_2 > 0$ such that for $\bar{t}_1 < t < \bar{t}_2$,*

$$\int (-\Psi'' + W''(\Theta)\Psi)\Psi\,\xi^2\,dz \geq \zeta_1 \int |\Psi'|^2\xi^2\,dz + \zeta_2 \int |\Psi|^2\xi^2\,dz - \frac{2}{\zeta_1}\int |\Psi|^2|\xi'|^2\,dz,$$

where integration is either over $(-\varepsilon^{-\beta} + z^{m_0}, z^{m_0})$, $(0, \varepsilon^{-\beta})$, or $(z^{m_0}, 0)$.

This proposition is very similar to the proof of Proposition 8 in [S1] and we include its proof in the appendix.

We are now left with estimating all of the terms in the right-hand side of (5.16). For this we find further estimates on $\Phi_\varepsilon - \Phi_0^\varepsilon$ and on Φ_1^ε. First, we present a bound on Φ_1^ε, which in particular gives the estimate $\|\Theta^\varepsilon - \Phi_0^\varepsilon\|_{H^{1,\infty}(z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon})\times(\bar{t}_1,\bar{t}_2)} \leq C\sqrt{\varepsilon}$.

LEMMA 5.7.

$$\|\Phi_1^\varepsilon\|_{H^{1,\infty}(z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon})\times(\bar{t}_1,\bar{t}_2)} \leq C\,|\lambda_\varepsilon(t)|\,.$$

The proof follows the same lines as the proof of Lemma 7 in [S1]. The basic tool is the following representation formula (cf. [BF]):

$$\Phi_{1i}^\varepsilon(z + z_\varepsilon^i, t) = \left(A(z)\int_0^z B + B(z)\int_z^\infty A\right)\lambda_\varepsilon(t),$$

where $A(z) := 1 - \tanh^2(z)$ and $B(z) := -A(z)\int_0^z \frac{1}{A^2}$.

As yet another consequence of Proposition 3.2, we have the following lemma, which implies that Φ is close to Φ_0 in the topology of L^∞.

LEMMA 5.8. *For any $\delta > 0$ there exist $e(\delta) > 0$ and $M(\delta) > 0$ such that*

$$\left|z_\varepsilon^1 - \frac{1 - r_\varepsilon^1}{\varepsilon}\right| \geq e(\delta) \quad \textit{and} \quad |z_\varepsilon^i - z_\varepsilon^{i-1}| \geq e(\delta) \quad \textit{for } 2 \leq i \leq m_0,$$

$$\|(\Psi^\varepsilon\Theta^\varepsilon)_-\|_{L^\infty((z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon})\times(\bar{t}_1,\bar{t}_2))} \leq M(\delta),$$

with $M(\delta) \to 0$ and $e(\delta) \to \infty$ as $\delta \to 0$.

This lemma is easily derived from explicit integration of the ODE (cf. [S1])

$$\begin{cases} |\Phi'(z,t)|^2 - 2W(\Phi(z,t)) = 2K_\varepsilon(z,t), \\ \Phi(z_i^\varepsilon, t) = a_\varepsilon \end{cases} \tag{5.19}$$

in $z_\varepsilon^- \leq z \leq z_\varepsilon^+$, where $\|K_\varepsilon\|_{L^\infty(z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon})\times(\bar{t}_1,\bar{t}_2)} \leq \delta^2$ by Proposition 3.2.

We are now ready for the proof of the first-order approximation Proposition 5.5.

Proof of Proposition 5.5. Let S be $(-\varepsilon^{-\beta} + z^{m_0}, z^{m_0})$, $(0, \varepsilon^{-\beta})$, or $(z^{m_0}, 0)$. We multiply equation (5.16) for Ψ by $\xi^2\Psi$, where ξ is defined by (5.10); then we integrate in z and t and use Proposition 5.6 to obtain

$$\begin{aligned} &\zeta_1 \int_{\bar{t}_1}^{\bar{t}_2}\int_S |\Psi'|^2\xi^2\,dz\,dt + \zeta_2 \int_{\bar{t}_1}^{\bar{t}_2}\int_S |\Psi|^2\xi^2\,dz\,dt \\ &\qquad \leq \frac{2}{\zeta_2}\int_{\bar{t}_1}^{\bar{t}_2}\int_S |F_\varepsilon|^2\xi^2 + |H_\varepsilon|^2\xi^2\,dz\,dt + \frac{2}{\zeta_1}\int_{\bar{t}_1}^{\bar{t}_2}\int_S |\Psi|^2|\xi'|^2\,dz\,dt \\ &\qquad\qquad + \left(\frac{\zeta_2}{4} + 6\|(\Psi\Theta)_-\|_{L^\infty}\right)\int_{\bar{t}_1}^{\bar{t}_2}\int_S |\Psi|^2\xi^2\,dz\,dt. \end{aligned} \tag{5.20}$$

Let $M(\delta)$ be given by Lemma 5.8 and choose δ small enough that $6M(\delta) \leq \frac{\zeta_2}{4}$. Incorporating this into (5.20) yields

$$\begin{aligned} &\zeta_1 \int_{\bar t_1}^{\bar t_2} \int_S |\Psi'|^2 \xi^2 \, dz \, dt + \frac{\zeta_2}{2} \int_{\bar t_1}^{\bar t_2} \int_S |\Psi|^2 \xi^2 \, dz \, dt \\ &\leq \frac{2}{\zeta_2} \int_{\bar t_1}^{\bar t_2} \int_S (|F_\varepsilon|^2 + |H_\varepsilon|^2) \xi^2 \, dz \, dt + \frac{2}{\zeta_1} \int_{\bar t_1}^{\bar t_2} \int_S |\Psi|^2 |\xi'|^2 \, dz \, dt. \end{aligned} \tag{5.21}$$

Thus we are left with estimating the right-hand side of (5.21). Using the definition of F_ε given by (5.13), the energy bound implies (if $\bar r_0 \neq 1$)

$$\int_{\bar t_1}^{\bar t_2} \int_{-\varepsilon^{-\beta}+z^{m_0}}^{\varepsilon^{-\beta}} |F_\varepsilon|^2 \, \xi^2 \, dz \, dt \leq C(R_0) \varepsilon^2. \tag{5.22}$$

Moreover, we have the estimate

$$\int_{\bar t_1}^{\bar t_2} \int_{(-\varepsilon^{-\beta}+z^{m_0}, z^{m_0}) \cup (0, \varepsilon^{-\beta})} |H_\varepsilon|^2 \xi^2 \, dz \, dt \leq C(R_0) \varepsilon^{3-\beta} \leq C_(R_0) \varepsilon^{2\beta}. \tag{5.23}$$

Indeed, the last two sums in the definition in (5.15) of H_ε drop out since in $(-\varepsilon^{-\beta} + z^{m_0}, z^{m_0})$, the function $\Xi_{m_0} \equiv 1$ and $\Xi_i \equiv 0$ for $1 \leq i \leq m_0 - 1$, while in $(0, \varepsilon^{-\beta})$, the functions $\Xi_i \equiv 0$ for $2 \leq i \leq m_0$ and $\Xi_1 \equiv 1$. Then (5.23) follows from the fact that $\|\Phi_0\|_\infty \leq 1$, Lemma 5.7, and the L^2 and L^∞ bounds on $\lambda_\varepsilon(t)$ (cf. Corollary 2.2 and Proposition 2.3).

In addition, we claim that

$$\int_{\bar t_1}^{\bar t_2} \int_{z^{m_0}}^{0} |H_\varepsilon|^2 \, dz \, dt \to 0 \tag{5.24}$$

as $\varepsilon \to 0$. To prove this, we need a better estimate on $|z^{i+1} - z^i|$ than the one we obtained in Lemma 5.8. If we go back to formula (3.6) in Proposition 3.1, we can easily obtain the following bound for almost all $t \in (\bar t_1, \bar t_2)$ by using the energy estimates (2.3) and (2.4) and Proposition 4.1 at the small expenditure of an $\ln(\frac{1}{\varepsilon})$ factor:

$$\| -\varepsilon^2 (\varphi'(\cdot, t))^2 + 2W(\varphi(\cdot, t)) \|_{L^\infty(R_0, 1)} \leq C(t, R_0) \varepsilon \ln \left(\frac{1}{\varepsilon} \right). \tag{5.25}$$

Therefore, in fact, we have for almost all $t \in (\bar t_1, \bar t_2)$ that

$$|z^{i+1} - z^i| \to \infty \quad \text{as } \varepsilon \to 0,$$

and therefore

$$|\Theta_{i+1} - \Theta_i| \to 0 \quad \text{in } \{\Xi_i \Xi_{i+1} \neq 0\}.$$

Since Θ_i are uniformly bounded, we can now conclude that $\int_{\bar t_1}^{\bar t_2} \int_{\{\Xi_i \Xi_{i+1} \neq 0\}} |\Theta_{i+1} - \Theta_i|^2 \, dz \, dt \to 0$ as $\varepsilon \to 0$, and hence (5.24) follows.

Finally, since $|\xi'(z)| \leq C\frac{1}{\varepsilon^{-\beta}-\varepsilon^{-\alpha}} < C\varepsilon^{\beta}$, the last term can be estimated as follows:

$$
\begin{aligned}
&\frac{2}{\zeta_1}\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\beta}+z^{m_0}}^{\varepsilon^{-\beta}} |\Psi|^2|\xi'|^2\,dz\,dt \leq C\varepsilon^{2\beta}\iint_{\{\xi\neq 0,1\}} |\Psi|^2\,dz\,dt \\
&\leq C\varepsilon^{2\beta}\iint_{\{\xi\neq 0,1\}} (|\Phi-\Phi_0|^2+\varepsilon^2|\Phi_1|^2)\,dz\,dt \\
&\leq C\frac{\varepsilon^{2\beta-1}}{R_0^{n-1}}\int_{\bar{t}_1}^{\bar{t}_2}\left(\int_{r_\varepsilon^i(t)}^{b+\eta}(\varphi(r,t)-1)^2 r^{n-1}\,dr + \int_{a+\eta}^{r_\varepsilon^{m_0}}(\varphi(r,t)+1)^2 r^{n-1}\,dr\right) dtC\varepsilon^{\beta+1},
\end{aligned}
$$

where we have used the fact that in $\{(z,t)|\xi(z,t)\neq 0,1\}$, the function Φ_0 is exponentially close to ± 1 depending on $z>0$ or $z<z^{m_0}$, while $\varepsilon^2\|\Phi_1\|^2_{L^\infty(z_\varepsilon^- - \frac{\eta}{\varepsilon}, z_\varepsilon^+ + \frac{\eta}{\varepsilon})\times(\bar{t}_1,\bar{t}_2)} < C\varepsilon$. Now using the energy bound (2.4), we obtain

$$
\frac{2}{\zeta_1}\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\beta}+z^{m_0}}^{\varepsilon^{-\beta}} |\Psi|^2|\xi'|^2\,dz\,dt \leq C(R_0)\left(\varepsilon^{2\beta}+\varepsilon^{\beta+1}\right) \leq C(R_0)\varepsilon^{2\beta}. \tag{5.26}
$$

This completes the proof for $\bar{r}_0 < 1$. If $\bar{r}_0 = 1$, all of the above arguments are valid except for those leading to (5.22) and (5.26). If in (5.22) we integrate over $(-\varepsilon^{-\beta} + z_\varepsilon^{m_0}, \frac{1-r_\varepsilon^1}{\varepsilon})$, the result is the same. For $z > \frac{1-r_\varepsilon^1}{\varepsilon}$, we have $F_\varepsilon = W'(\varphi_\varepsilon(t,1)) - \varepsilon\lambda_\varepsilon(t)$. However, with (5.25) and because $\varphi_\varepsilon'(1,t) = 0$, we find $F_\varepsilon^2 \leq C(t)\varepsilon ln(1/\varepsilon)$, and the remaining integral in (5.22) still converges to zero by Lebesgue's convergence theorem. The same reasoning holds for (5.26). □

Finally, we conclude with some $H^{1,\infty}$ bounds from this $H^{1,2}$ bound.

COROLLARY 5.9. *Let* $J = J_{\text{good}}(t) := (-\varepsilon^{-\alpha} + z^{m_0}, z^{m_0}) \cup (0, \varepsilon^{-\alpha})$ *if* $\bar{r}_0 \neq 1$ *and* $J = J_{\text{good}}(t) := (-\varepsilon^{-\alpha} + z^{m_0}, z^{m_0})$ *if* $\bar{r}_0 = 1$. *Then*

$$
\begin{aligned}
&\int_{\bar{t}_1}^{\bar{t}_2} \|\Psi_\varepsilon'\|^2_{H^{1,\infty}(J)}\,dt \leq C\varepsilon^{2\beta}, \\
&\int_{\bar{t}_1}^{\bar{t}_2} \|\Psi_\varepsilon'\|^2_{H^{1,\infty}(0,z^{m_0})}\,dt \to 0, \\
&\int_{\bar{t}_1}^{\bar{t}_2} \|\Phi^\varepsilon(\cdot,t)-\Phi_0^\varepsilon(\cdot,t)\|^2_{H^{1,2}(-\varepsilon^{-\alpha}+z^{m_0},\varepsilon^{-\alpha})}\,dt \to 0, \\
&\int_{\bar{t}_1}^{\bar{t}_2} \|\Phi^\varepsilon(\cdot,t)-\Phi_0^\varepsilon(\cdot,t)\|^2_{H^{1,\infty}(J)}\,dt \leq C\varepsilon^{2\beta}, \\
&\int_{\bar{t}_1}^{\bar{t}_2} \|\Phi^\varepsilon(\cdot,t)-\Phi_0^\varepsilon(\cdot,t)\|^2_{H^{1,\infty}(0,z^{m_0})}\,dt \to 0,
\end{aligned}
$$

where C *depends on the data and* B.

Proof. The first two results follow from Proposition 5.5 by the Sobolev embedding theorem in $\mathbf{R}$ since Ψ satisfies the differential equation (5.16) and thus Ψ'' satisfies the same bounds as Ψ and Ψ'.

The last three results follow from either Proposition 5.5 or the two first results of this corollary using the by now "familiar" estimates

$$
\int_{\bar{t}_1}^{\bar{t}_2} \|\varepsilon\Phi_1^\varepsilon(\cdot,t)\|^2_{H^{1,2}(-\varepsilon^{-\alpha},z^{m_0}+\varepsilon^{-\alpha})}\,dt \leq \varepsilon^2\frac{1}{\varepsilon}\int_{\bar{t}_1}^{\bar{t}_2} |\lambda_\varepsilon(t)|^2\,dt \leq C\varepsilon
$$

and

$$\int_{\bar t_1}^{\bar t_2} \|\varepsilon\Phi_1^\varepsilon(\cdot,t)\|^2_{H^{1,\infty}(-\varepsilon^{-\alpha},z^{m_0}+\varepsilon^{-\alpha})}\,dt \le \varepsilon^2\int_{\bar t_1}^{\bar t_2}|\lambda_\varepsilon(t)|^2\,dt \le C\varepsilon^2. \qquad \square$$

6. The limit equation. In this section, we derive the differential equation of the limiting interface (true or phantom) given by $r=\bar r^i$. Therefore, we once again assume (3.1)–(3.5) for φ_ε, and hence the analysis of the preceding sections applies.

PROPOSITION 6.1. *Let A_{R_0} be as in* (4.2). *Then for all $(\bar t_0,\bar r_0)\in\Gamma_{R_0}$ as in* (4.6), *there exists a box $B=(\bar t_1,\bar t_2)\times(a,b)$ as in Proposition* 4.2 *such that $(\bar r_0,\bar t_0)\in B$ and such that $\Gamma_{R_0}\cap B$ consists of $m_0=m_0(\bar r_0,\bar t_0)$ Hölder-$\frac12$ graphs $\bar r^i:(\bar t_1,\bar t_2)\to(a,b)$ with $\bar r^i(\bar t_0)=\bar r_0$ (where $i=k,\dots,m_0+k-1$). Moreover, if $\bar r_0\neq 1$,*

$$-\sum_{i=k}^{m_0+k-1}\left(\dot{\bar r}^i+\frac{n-1}{\bar r^i}\right)=\frac32\nu(\bar r_0,\bar t_0)\lambda_0 \qquad \textit{in } (\bar t_1,\bar t_2),$$

and $\le$ is true in the above if $\bar r_0=1$.

The constant $m_0(\bar r_0,\bar t_0)$ can be understood as the "multiplicity" of Γ_{R_0}. We shall prove that $m_0=1$ almost everywhere in $\{\nu\neq0\}$, i.e., that the multiplicity of true interfaces is 1. For this we show later that $m_0(\bar r_0,\bar t_0)\neq1$ in B is equivalent to $\nu(\bar r_0,\bar t_0)\lambda_0(\bar t_0)=0$ (cf. Proposition 6.3). Then, using the fact that the mass is preserved in time, we calculate explicitly λ_0 and show that $\lambda_0(\bar t_0)\neq0$ (cf. Proposition 6.5).

Proof of Proposition 6.1. For the simplicity of presentation we assume $k=1$. We first consider the case $\bar r_0\neq1$ and later point out the differences in the other case.

In the z-variable the nonlocal equation (3.1) becomes

$$\varepsilon\partial_t\Phi-\dot r^1\Phi'-\frac1\varepsilon\Phi''-\frac{n-1}{\varepsilon z+r^1}\Phi'+\frac1\varepsilon W'(\Phi)-\lambda_\varepsilon(t)=0. \tag{6.1}$$

Let ζ be a smooth time-dependent test function with compact support in $(\bar t_1,\bar t_2)$. First, we multiply equation (6.1) by $\Phi'\zeta$ and integrate over $(-\varepsilon^{-\alpha}+z^{m_0},\varepsilon^{-\alpha})\times(\bar t_1,\bar t_2)$,

$$\begin{aligned}\int_{\bar t_1}^{\bar t_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\zeta(\varepsilon\Phi_t\Phi'-\dot r^1\Phi'^2)\,dz\,dt-\frac1\varepsilon\int_{\bar t_1}^{\bar t_2}\zeta\left(\frac12(\Phi')^2-W(\Phi)\right)\Bigg|_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}dt\\ =\int_{\bar t_1}^{\bar t_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\frac{(n-1)\zeta}{\varepsilon z+r^1}|\Phi'|^2\,dz\,dt+\int_{\bar t_1}^{\bar t_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\zeta\lambda_\varepsilon\Phi'\,dz\,dt.\end{aligned} \tag{6.2}$$

The main part of this section is the evaluation of the limit of each term in (6.2) as $\varepsilon\to0$.

Second term. We claim that

$$\int_{\bar t_1}^{\bar t_2}\frac1\varepsilon\left(\frac12(\Phi')^2-W(\Phi)\right)\Bigg|_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\zeta\,dt\overset{\varepsilon\to0}{\to}0. \tag{6.3}$$

Indeed, the endpoints $-\varepsilon^{-\alpha}+z^{m_0}$ and $\varepsilon^{-\alpha}$ are in J_{good}, so we can apply Corollary 5.9 and replace Φ by Φ_0 since $\int_{\bar t_1}^{\bar t_2}\|\Phi-\Phi_0\|^2_{H^{1,\infty}(J)}\,dt\le C\varepsilon^{2\beta}$ and $2\beta-1>0$. The claim in (6.3) now follows since $\Phi_0(-\varepsilon^{-\alpha}+z^{m_0},t)=\tanh(\pm\varepsilon^{-\alpha}+\mu^\varepsilon)$ and $\Phi_0(\varepsilon^{-\alpha},t)=\tanh(\varepsilon^{-\alpha}+\mu^\varepsilon)$.

Third term. We claim that

$$(n-1)\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta\,\frac{1}{\varepsilon z + r_\varepsilon^1}|\Phi'|^2\,dz\,dt \overset{\varepsilon\to 0}{\to} \frac{4}{3}(n-1)\int_{\bar{t}_1}^{\bar{t}_2}\zeta\sum_{i=1}^{m_0}\frac{1}{\bar{r}^i}\,dt. \tag{6.4}$$

To prove this, we first replace Φ by Φ_0. We can do this since $|\frac{1}{\varepsilon z + r_\varepsilon^1}| \leq \frac{1}{R_0}$ and since by the approximation proposition (Proposition 5.5), Φ' is well approximated by Φ_0'.

Therefore, we only have to consider the limit of the integral

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta\frac{1}{\varepsilon z + r_\varepsilon^1}|\Phi_0'|^2\,dz\,dt. \tag{6.5}$$

To find this limit, we divide the interval of integration in subintervals each of which contain one interface:

$$\begin{aligned}(-\varepsilon^{-\alpha}+z^{m_0},\varepsilon^{-\alpha}) &= C_{m_0} \cup \bigcup_{i=2}^{m_0-1} C_i \cup C_1 \cup \bigcup_{i=1}^{m_0} B_i \\ &:= \left(-\varepsilon^{-\alpha}+z^{m_0}, \frac{z^{m_0}+z^{m_0-1}}{2}-1\right) \\ &\cup \bigcup_{i=2}^{m_0-1}\left(\frac{z^{i+1}+z^i}{2}+1, \frac{z^i+z^{i-1}}{2}-1\right) \\ &\cup \left(\frac{z^2+z^1}{2}+1, \varepsilon^{-\alpha}\right) \cup \bigcup_{i=1}^{m_0}\{\Xi_i \neq 0,1\}.\end{aligned} \tag{6.6}$$

The set $\{\Xi_i \neq 0,1\}$ is of length 4, we have the estimate $|\frac{1}{\varepsilon z + r_\varepsilon^1}| \leq \frac{1}{R_0}$, and we know that $\|\Phi_0'(z)\|_{L^\infty(\{\Xi_i\neq 0,1\})} \overset{\varepsilon\to 0}{\to} 0$ almost everywhere since by (5.25), $|z^{i+1}-z^i| \overset{\varepsilon\to 0}{\to} \infty$. Thus the integral over B_i gives 0 in the limit.

We note that by the definition in (5.6) of Φ_0, we have $\Phi_0(z+z^i) = \tanh((-1)^{i+1}z+\mu_\varepsilon)$ for $z \in C_i - z^i$; therefore, we claim that

$$\begin{aligned}&\sum_{i=1}^{m_0}\int_{\bar{t}_1}^{\bar{t}_2}\int_{C_i-z^i}\frac{\zeta}{\varepsilon z + r_\varepsilon^i}{\Phi_0'}^2(z+z^i)\,dz\,dt \\ &\overset{\varepsilon\to 0}{\to} \sum_{i=1}^{m_0}\int_{\bar{t}_1}^{\bar{t}_2}\frac{\zeta}{\bar{r}^i}\int_{-\infty}^{\infty}(1-\tanh^2(z))^2\,dz\,dt = \frac{4}{3}\int_{\bar{t}_1}^{\bar{t}_2}\zeta\sum_{i=1}^{m_0}\frac{1}{\bar{r}^i}\,dt\end{aligned} \tag{6.7}$$

since $r_\varepsilon^i \to \bar{r}^i$ uniformly (cf. Proposition 4.2) and since $1-\tanh^2$ decays exponentially.

Last term. We claim that

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\zeta\,\lambda_\varepsilon\Phi'\,dz\,dt \overset{\varepsilon\to 0}{\to} 2\nu\int_{\bar{t}_1}^{\bar{t}_2}\zeta\,\lambda_0\,dt, \tag{6.8}$$

where ν was introduced in Proposition 4.2 and λ_0 was introduced in Remark 2.4. Indeed, integrating by parts yields

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\zeta\,\lambda_\varepsilon\Phi'\,dz\,dt = \int_{\bar{t}_1}^{\bar{t}_2}\zeta\,\lambda_\varepsilon\ \Phi|_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}}\ dt.$$

However,

$$\Phi(\varepsilon^{-\alpha},t) \quad \text{and} \quad \Phi(-\varepsilon^{a}+z^{m_0},t) \longrightarrow \pm 1$$

by the approximation proposition (Proposition 5.5), and since $\lambda_\varepsilon \longrightarrow \lambda_0$ in L^2, the claim in (6.8) is immediate. Note that we allow for $\nu = \{-1,0,1\}$.

First term. We claim that

$$\int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta(\varepsilon \Phi_t \Phi' - \dot{r}^1 {\Phi'}^2)\, dz\, dt \overset{\varepsilon \to 0}{\to} -\int_{\bar{t}_1}^{\bar{t}_2} \frac{4}{3} \zeta \sum_{i=1}^{m_0} \partial_t \bar{r}^i\, dt. \tag{6.9}$$

This proof is similar to the convergence of the fourth term. The fourth term splits into two distinctly different parts:

$$\begin{aligned} &\int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta \Phi'(\varepsilon \Phi_t - \dot{r}^1 \Phi')\, dz\, dt \\ &\qquad = \int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \varepsilon \zeta \Phi' \Phi_t\, dz\, dt - \int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta \dot{r}^1 {\Phi'}^2\, dz\, dt. \end{aligned} \tag{6.10}$$

The last term should describe the dynamics in the limit, whereas the first term will vanish. This is the rigorous verification of the traveling-wave structure of the solution. We claim that

$$\begin{aligned} &\int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \zeta \dot{r}^1 {\Phi'}^2\, dz\, dt \\ &\qquad = \sum_{i=1}^{m_0} \int_{\bar{t}_1}^{\bar{t}_2} \int_{[C_i \cup B_i]-z^i} \zeta \dot{r}^i (\Phi'(z+z^i,t))^2\, dz\, dt \overset{\varepsilon \to 0}{\to} \int_{\bar{t}_1}^{\bar{t}_2} \frac{4}{3} \zeta \sum_{i=1}^{m_0} \partial_t \bar{r}^i\, dt. \end{aligned} \tag{6.11}$$

We know from Proposition 4.2 that $\partial_t r^i \longrightarrow \partial_t \bar{r}^i$ weakly in $L^2(\bar{t}_1,\bar{t}_2)$. Therefore, (6.11) is true if we show that

$$\int_{[C_i \cup B_i]-z^i} (\Phi'(z+z^i,t))^2\, dz \longrightarrow \frac{4}{3} \quad \text{strongly in } L^2(\bar{t}_1,\bar{t}_2). \tag{6.12}$$

This amounts to showing that we can replace Φ by Φ_0 in (6.11) since $\int_{-\infty}^{\infty} (\Phi_0')^2\, dz = \int_{-\infty}^{\infty} (1-\tanh^2(z))^2\, dz = \frac{4}{3}$, but this follows from Corollary 5.9 and the energy bound.

Next, we claim that

$$\begin{aligned} &\int_{\bar{t}_1}^{\bar{t}_2} \int_{-\varepsilon^{-\alpha}+z^{m_0}}^{\varepsilon^{-\alpha}} \varepsilon \zeta \Phi' \Phi_t\, dz\, dt \\ &\qquad = \sum_{i=1}^{m_0} \int_{\bar{t}_1}^{\bar{t}_2} \int_{[C_i \cup B_i]-z^i} \varepsilon \zeta \Phi'(z+z^i,t) \Phi_t(z+z^i,t)\, dz\, dt \overset{\varepsilon \to 0}{\to} 0. \end{aligned} \tag{6.13}$$

First, as usual, we replace Φ' by Φ_0'. This time this is possible because

$$\varepsilon \left(\int (\Phi_t)^2\, dz\, dt \right)^{\frac{1}{2}} \leq C, \tag{6.14}$$

while $\int(\Phi' - \Phi_0')^2\,dz\,dt \longrightarrow 0$ by Corollary 5.9. In order to show estimate (6.14), we proceed as in [S1] and prove the following a priori estimate on $\partial_t\Phi$:

$$\begin{aligned}\int |\Phi_t|^2\,dz\,dt &\le 2\int |\Phi_t - \frac{\dot r^1}{\varepsilon}\Phi'|^2\,dz\,dt + \frac{2}{\varepsilon^2}\int |\dot r^1|^2|\Phi'|^2\,dz\,dt\\ &\le \frac{2}{\varepsilon R_0^{n-1}}\int_0^T\int_\Omega |\partial_t\varphi(x,t)|^2\,dx\,dt + \frac{2}{\varepsilon R_0^{n-1}}\\ &\underset{t\in(0,T)}{\rightarrow} \sup\int_\Omega |\nabla\varphi(x,t)|^2\,dx\int_0^T |\dot r^1|^2\,dt \le \frac{C(R_0)}{\varepsilon^2}.\end{aligned}$$

Therefore, we may substitute Φ' by Φ_0' in (6.13) and only have to show that

$$\varepsilon\sum_{i=1}^{m_0}\int_{\bar t_1}^{\bar t_2}\int_{[C_i\cup B_i]-z^i}\zeta\Phi_t\Phi_0'(z+z^i,t)\,dz\,dt \overset{\varepsilon\to 0}{\to} 0. \tag{6.15}$$

Since in B_i we know that $\Phi_0' \overset{\varepsilon\to0}{\to} 0$ pointwise, using (6.14), we obtain

$$\left|\int_{\bar t_1}^{\bar t_2}\int_{B_i-z^i}\varepsilon\zeta\Phi_t\Phi_0'(z+z^i,t)\right|\,dz\,dt \le C + \int_{\bar t_1}^{\bar t_2}\int_{B_i}(\Phi_0')^2\,dz\,dt \overset{\varepsilon\to0}{\to} 0.$$

Now we note the essential fact that in $C_i - z^i$ we have $\Phi_0(z+z^i) = \tanh((-1)^{i+1}z+\mu_\varepsilon)$, so $\partial_t\Phi_0' = 0$. Therefore, we obtain

$$\begin{aligned}\varepsilon\sum_{i=1}^{m_0}\int_{\bar t_1}^{\bar t_2}\int_{C_i-z^i}\zeta\Phi_t\Phi_0'(z+z^i,t)\,dz\,dt &= -\varepsilon\sum_{i=1}^{m_0}\int_{\bar t_1}^{\bar t_2}\int_{B_i-z^i}\partial_t\zeta\Phi\Phi_0'(z+z^i,t)\,dz\,dt\\ - \varepsilon\sum_{i=1}^{m_0}\int_{\bar t_1}^{\bar t_2}\zeta\Phi\Phi_0'|_{\frac{z^{i+1}+z^i}{2}+1}\frac{\dot r^{i+1}-\dot r^i}{2\varepsilon}\,dt &+ \varepsilon\sum_{i=1}^{m_0}\int_{\bar t_1}^{\bar t_2}\zeta\Phi\Phi_0'|_{\frac{z^{i-1}+z^i}{2}-1}\frac{\dot r^{i-1}-\dot r^i}{2\varepsilon}\,dt.\end{aligned}$$

Since $\Phi_0'(\cdot + z^i) \in L^1(R)$ and since Φ is bounded, the first sum converges to 0 with ε. For the second and third terms, we use the fact that $\dot r^i$ converge weakly in L^2 and the fact that $\Phi_0'(\frac{z^{i+1}+z^i}{2} - 1) \overset{\varepsilon\to0}{\to} 0$ pointwise and thus in L^2. This finishes the proof of (6.15) and hence of (6.9).

If $\bar r_0 = 1$, the only difference in our strategy is that we integrate (6.1) over $(-\varepsilon^{-\alpha} + z^{m_0}, \min(\varepsilon^{-\alpha}, \frac{1-r_\varepsilon^1}{\varepsilon}))$. Then everything remains the same, only (6.3) must be changed into

$$\lim_{\varepsilon\to0}\int_{\bar t_1}^{\bar t_2}\frac{1}{\varepsilon}\left(\frac{1}{2}(\Phi')^2 - W(\Phi)\right)\Bigg|_{-\varepsilon^{-\alpha}+z^{m_0}}^{\min(\varepsilon^{-\alpha},\frac{1-r_\varepsilon^1}{\varepsilon})}\zeta\,dt \le 0.$$

In summary, we have evaluated the limit of each term in (6.2). Therefore, using equation (6.2) and the limit of each term given by (6.3), (6.4), (6.8), and (6.9), we have shown that in the limit as $\varepsilon \to 0$, we obtain the equation

$$-\int_{\bar t_1}^{\bar t_2}\frac{4}{3}\zeta\sum_{i=1}^{m_0}\left(\dot{\bar r}^i + \frac{n-1}{\bar r^i}\right)dt = 2\nu\int_{\bar t_1}^{\bar t_2}\zeta\lambda_0\,dt, \tag{6.16}$$

or the inequality with $\le$ if $\bar r_0$ was in the fixed boundary of the domain. This limit was obtained in the box B defined in Proposition 4.2. □

The next lemma is essential in studying the issue of multiplicities.

LEMMA 6.2. *Let $(\bar{t}_0, \bar{r}_0)$ be as in Proposition* 6.1. *Then if $\nu(\bar{t}_0, \bar{r}_0) \neq 0$ and $\bar{r}_0 \neq 1$,*

$$-\left(\dot{\bar{r}}^k + \frac{n-1}{\bar{r}^k}\right) \geq \frac{3}{2}\nu(\bar{r}_0, \bar{t}_0)\lambda_0 \geq -\left(\dot{\bar{r}}^{m_0+k-1} + \frac{n-1}{\bar{r}^{m_0+k-1}}\right) \qquad \textit{in } (\bar{t}_1, \bar{t}_2), \tag{6.17}$$

and if $\bar{r}_0 = 1$, then the second of the above inequalities holds.

Proof. We again assume that $k = 1$. Without loss of generality, assume that $\nu = 1$. Now, instead of integrating (6.2) over $(-\varepsilon^{-\alpha} + z^{m_0}, \varepsilon^{-\alpha})$, we integrate only over the last branch of Φ. More precisely, let $c_\varepsilon(t)$ be the first negative point where $\Phi'(\cdot, t)$ vanishes. We note that in the same way that we proved that $|z^{i+1} - z^i| \overset{\varepsilon\to 0}{\to} \infty$ (cf. 5.25), we can show that $c_\varepsilon(t) \overset{\varepsilon\to 0}{\to} -\infty$ pointwise. We integrate (6.2) over $(c_\varepsilon, \varepsilon^{-\alpha})$ and use the same arguments as before to obtain the limit as $\varepsilon \to 0$. This gives an inequality for $\dot{\bar{r}}_1$ similar to (6.16). We only point out the main differences with the previous case. The claim in (6.3) is replaced by

$$\lim_{\varepsilon\to 0}\int_{\bar{t}_1}^{\bar{t}_2} \zeta\frac{1}{\varepsilon}\left(\frac{1}{2}(\Phi')^2 - W(\Phi)\right)\Bigg|_{c_\varepsilon}^{\varepsilon^{-\alpha}} dt = \int_{\bar{t}_1}^{\bar{t}_2} \zeta \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}W(\Phi(c_\varepsilon))\, dt,$$

and the proof is the same as before since $\Phi'(c_\varepsilon(t), t) = 0$.

The claim in (6.4) is replaced by

$$(n-1)\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}}^{c_\varepsilon} \zeta\frac{1}{\varepsilon z + r^1}|\Phi'|^2\, dz\, dt \overset{\varepsilon\to 0}{\to} \frac{4}{3}(n-1)\int_{\bar{t}_1}^{\bar{t}_2}\zeta\frac{1}{\bar{r}^1}\, dt,$$

and the proof is similar to that above since $c_\varepsilon \overset{\varepsilon\to 0}{\to} -\infty$.

The claim (6.8) is replaced by

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{-\varepsilon^{-\alpha}}^{c_\varepsilon} \zeta\lambda_\varepsilon\Phi'\, dz\, dt \overset{\varepsilon\to 0}{\to} 2\int_{\bar{t}_1}^{\bar{t}_2}\zeta\lambda_0\, dt.$$

The proof must be modified slightly since we only know that $\Phi(c_\varepsilon, t) \to -1$ pointwise and thus in L^2. However, this is enough to pass to the limit.

The claim (6.9) is replaced by

$$\int_{\bar{t}_1}^{\bar{t}_2}\int_{c_\varepsilon}^{\varepsilon^{-\alpha}} \zeta(\varepsilon\Phi_t\Phi' - \dot{r}^1{\Phi'}^2)\, dz\, dt \overset{\varepsilon\to 0}{\to} -\int_{\bar{t}_1}^{\bar{t}_2}\frac{4}{3}\zeta\partial_t\bar{r}^1\, dt,$$

and the proof is the same.

Thus, in the limit, we find that

$$-\int_{\bar{t}_1}^{\bar{t}_2}\frac{4}{3}\zeta\left(\dot{\bar{r}}^1 + \frac{n-1}{\bar{r}^1}\right)dt - \int_{\bar{t}_1}^{\bar{t}_2}\zeta\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}W(\Phi(c_\varepsilon))\, dt = 2\int_{\bar{t}_1}^{\bar{t}_2}\zeta\lambda_0\, dt.$$

Similarly, we can integrate (6.2) over $(-\varepsilon^{-\alpha} + z^{m_0}, d_\varepsilon)$, where d_ε is the first point to the right of z^{m_0} for which Φ' vanishes. Since $\Phi_0(-\varepsilon^{-\alpha} + z^{m_0}) \overset{\varepsilon\to 0}{\to} -1$ and $\Phi_0(d_\varepsilon) \overset{\varepsilon\to 0}{\to} 1$, this yields

$$-\int_{\bar{t}_1}^{\bar{t}_2}\frac{4}{3}\zeta\left(\dot{\bar{r}}^{m_0} + \frac{n-1}{\bar{r}^{m_0}}\right)dt + \int_{\bar{t}_1}^{\bar{t}_2}\zeta\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}W(\Phi(d_\varepsilon))\, dt = 2\int_{\bar{t}_1}^{\bar{t}_2}\zeta\lambda_0\, dt$$

in the limit. Since $W \geq 0$, we conclude for $\zeta \geq 0$ that

$$-\int_{\bar{t}_1}^{\bar{t}_2} \frac{4}{3}\zeta\left(\dot{\bar{r}}^1 + \frac{n-1}{\bar{r}^1}\right) dt \geq 2\int_{\bar{t}_1}^{\bar{t}_2} \zeta\lambda_0 \, dt,$$
$$-\int_{\bar{t}_1}^{\bar{t}_2} \frac{4}{3}\zeta\left(\dot{\bar{r}}^{m_0} + \frac{n-1}{\bar{r}^{m_0}}\right) dt \leq 2\int_{\bar{t}_1}^{\bar{t}_2} \zeta\lambda_0 \, dt.$$

This proves the lemma. □

Proposition 6.3 (evolution equation). *Let A_{R_0} be as in* (4.2). *Then the interfaces $\bar{r} = \bar{r}^i$ defined in* (4.5) *evolve in their domain of existence $I^i_{R_0} \cap \{\bar{r}^i < 1\}$ according to*

$$-\left(\dot{\bar{r}} + \frac{n-1}{\bar{r}}\right) = \frac{3}{2}\nu(\cdot,\bar{r})\lambda_0. \tag{6.18}$$

In addition, for almost all $(\bar{t}_0, \bar{r}_0) \in \Gamma_{R_0}$ with $\bar{r}_0 < 1$,

$$\textit{either} \quad \nu(\bar{t}_0,\bar{r}_0)\lambda_0(\bar{t}_0) = 0 \quad \textit{or} \quad m_0(\bar{t}_0,\bar{r}_0) = 1. \tag{6.19}$$

Proof. Define for any $M_0 \geq k \geq j \geq 1$ the set (formally setting $\bar{r}^0 = +\infty$ and $\bar{r}^{M_0+1} = -\infty$)

$$I := \{\bar{t}_0 \in A_{R_0} \mid \bar{r}^{j-1} > \bar{r}^j = \cdots = \bar{r}^k > \bar{r}^{k+1} \text{ at } \bar{t}_0\}.$$

There exist only finitely many j and k as above, and any t has to be in one of the corresponding I's. Therefore, if we prove the claim in I, this proves the proposition.

For any $\bar{t}_0 \in I$ and $\bar{r}_0 := \bar{r}^j(\bar{t}_0)$,

$$m_0(\bar{t}_0,\bar{r}_0) = k - j + 1 \quad \text{and} \quad \nu(\bar{t}_0,\bar{r}_0) = \pm 1 \quad \text{if } m_0 \text{ is odd.}$$

Note that almost everywhere in I,

$$\dot{\bar{r}}^j = \cdots = \dot{\bar{r}}^k,$$

and thus Proposition 6.1 implies that almost everywhere in I (if $\bar{r}^j < 1$),

$$-m_0\left(\dot{\bar{r}}^j + \frac{n-1}{\bar{r}^j}\right) = -\sum_{i=j}^{k}\left(\dot{\bar{r}}^i + \frac{n-1}{\bar{r}^i}\right) = \frac{3}{2}\nu(\cdot,\bar{r}^j)\lambda_0. \tag{6.20}$$

If m_0 is even, then $\nu = 0$ almost everywhere in I and thus the differential equation (6.18) is immediate.

If m_0 is odd, then Lemma 6.2 applies, giving

$$-\left(\dot{\bar{r}}^j + \frac{n-1}{\bar{r}^j}\right) \geq \frac{3}{2}\nu(\cdot,\bar{r}^j)\lambda_0 \geq -\left(\dot{\bar{r}}^k + \frac{n-1}{\bar{r}^k}\right) \quad \text{in } I.$$

However, since both sides agree, the differential equation (6.18) is satisfied.

Finally, subtracting the differential equation (6.18) from (6.20) implies

$$(1 - m_0)\left(\dot{\bar{r}}^j + \frac{n-1}{\bar{r}^j}\right) = 0.$$

Thus either $m_0 = 1$ or $\dot{\bar{r}}^j + \frac{n-1}{\bar{r}^j} = 0$, the latter in turn implying that $\lambda_0 = 0$. This completes the proof. □

Now we have to identify the limit Lagrange multiplier λ_0. To determine λ_0 in terms of the interfaces of v, we use the mass-conservation property of nonlocal flow, namely

$$\partial_t \int_\Omega \varphi_\varepsilon(x,t)\, dx = 0 \quad \text{and hence} \quad \partial_t \int_\Omega v(x,t)\, dx = 0.$$

Since these are nonlocal quantities, we have to remove the constraint $r > R_0 > 0$.

Definition 6.4. *In* (4.5) *we defined interfaces*

$$\bar{r}^i : I^i_{R_0} \subset A_{R_0} \to (R_0, 1] \qquad \textit{for } i = 1, \ldots, M_0$$

with the property that $\bar{r}^i \geq \bar{r}^{i+1}$ and $\bar{r}^i = R_0$ on $\partial I^i_{R_0} \cap A_{R_0}$. These interfaces are locally Hölder-$\frac{1}{2}$.

For a sequence $R_0 \to 0$, we now define

$$A := \bigcap_{R_0 \to 0} A_{R_0}. \tag{6.21}$$

Then since this is a countable intersection, the complement of A has at most countably many points. In addition, we define $I^i := \bigcup_{R_0 \to 0} I^i_{R_0}$ and

$$R^i : I^i \longrightarrow (0,1] \qquad \textit{for } i = 1, \ldots \tag{6.22}$$

by $R^i(t) := \bar{r}^i(t)$ if $t \in I^i_{R_0}$. Then, in particular, $I^i_{R_0} \cap A = \{t \in A \mid R^i(t) > R_0\}$.

This definition is reasonable since on the intersections of $I^i_{R_0}$'s with different R_0, the corresponding $\bar{r}^i$'s agree. For any R_0, the $\bar{r}^i$'s were constructed (see section 4) by selecting subsequences, and using a diagonal sequence, we may assume that the subsequence does not depend on R_0. Thus the limit graphs $\bar{r}^i$ coincide in the intersections of the sets $A_{R_0} \times (R_0, 1]$.

We note that by definition, $m_0(\bar{t}_0, \bar{r}_0) = \#\{i \mid R^i(\bar{t}_0) = \bar{r}_0\}$ and $\nu_0(\bar{t}_0, \bar{r}_0) = 0$ if m_0 is even.

Proposition 6.5 (the Lagrange multiplier). *In the set A, the limit Lagrange multiplier is given by*

$$\lambda_0 = -\frac{2}{3} \frac{(n-1) \sum \nu(R^i)(R^i)^{n-2}}{\sum |\nu(R^i)|(R^i)^{n-1}},$$

where summation is over all interfaces of radii strictly less than 1. *If the space dimension is strictly larger than* 2, *then $\lambda_0 \neq 0$. If the space dimension is equal to* 2, *then $\lambda_0 \neq 0$ in case of an odd number of interfaces and identically zero in case of an even number of interfaces.*

Proof. Since $\varphi_\varepsilon \to v$ in L^1 and $v = \pm 1$ almost everywhere, the same is true for

$$\chi_\varepsilon := \frac{\varphi_\varepsilon - a_\varepsilon}{|\varphi_\varepsilon - a_\varepsilon|}.$$

By construction, for $t \in A_{R_0}$, the function $\chi_\varepsilon(t, \cdot)$ has discontinuities at $r^i_\varepsilon(t)$ in $[R_0, 1)$ for $i = 1, \ldots, M_0$ (formally setting $r^i_\varepsilon(t) = R_0$ if $r^i_\varepsilon(t)$ is not defined). Thus for $t \in A_{R_0}$

and $\varepsilon \to 0$,

$$\int_{|x|>R_0} v\,dx \longleftarrow \int_{|x|>R_0} \chi_\varepsilon\,dx = \pm\frac{1}{n}\left\{2\sum_{i=1}^{M_0}(-1)^{i+1}(r_\varepsilon^i)^n + (-1)^{M_0+1}(R_0)^n - 1\right\}$$
$$\longrightarrow \pm\frac{1}{n}\left\{2\sum_{i=1}^{M_0}(-1)^{i+1}(\bar{r}^i)^n + (-1)^{M_0+1}(R_0)^n - 1\right\},$$

where we formally set $\bar{r}^i(t) = R_0$ if $t \notin I^i_{R_0}$. Differentiating the resulting identity yields

$$\partial_t \int_{|x|>R_0} v\,dx = \pm 2\sum_{i=1}^{M_0}(-1)^{i+1}(\bar{r}^i)^{n-1}\dot{\bar{r}}^i \quad \text{for } t \in A_{R_0}.$$

This can be rewritten as

$$\partial_t \int_{|x|>R_0} v\,dx = \left\{\pm 2\sum_{i=1}^{M_0}(-1)^{i+1}(\bar{r}^i)^{n-1}\dot{\bar{r}}^i\right\}\chi + \mu_{R_0}, \tag{6.23}$$

where χ is the characteristic function of A and μ_{R_0} is a measure with support in $(0,T)\backslash A$.

Since $v \in BV((0,T)\times\Omega)$, we know that $\int_{|x|>R_0} v\,dx \in BV(0,T)$ uniformly bounded with respect to R_0 and thus, since $\int_{|x|>R_0} v\,dx \to \int_\Omega v\,dx$,

$$\partial_t \int_{|x|>R_0} v\,dx \overset{*}{\longrightarrow} 0 = \partial_t \int_\Omega v\,dx \quad \text{in } [C^0(0,T)]'. \tag{6.24}$$

We now use the differential equation (6.18) to calculate the limit of the sum in (6.23):

$$\sum_{i=1}^{M_0}(-1)^{i+1}(\bar{r}^i)^{n-1}\dot{\bar{r}}^i$$
$$= -\frac{3}{2}\lambda_0 \sum_{i=M_1(t)}^{M_2(t)}(-1)^{i+1}\nu(\bar{r}^i)(\bar{r}^i)^{n-1} - (n-1)\sum_{i=M_1(t)}^{M_2(t)}(-1)^{i+1}(\bar{r}^i)^{n-2},$$

where $M_2(t) := \#\{i \mid \bar{r}^i(t) > R_0\}$ and $M_1(t)$ is the first index such that the corresponding interface is less than 1.

We assume without loss of generality that v is positive near $\partial\Omega$. Then we remark that if $\nu(\bar{r}^i) \neq 0$, then $(-1)^{i+1} = -\nu(\bar{r}^i)$ because for any i with $\nu(\bar{r}^i) = 0$, an even number of interfaces collide. In the second sum, we may substitute $(-1)^{i+1}$ with $-\nu(\bar{r}^i)$ because either they agree or $(\bar{r}^i)^{n-2}$ is added up an even number of times with alternating signs. Thus

$$\sum_{i=1}^{M_0}(-1)^{i+1}(\bar{r}^i)^{n-1}\dot{\bar{r}}^i$$
$$= -\frac{3}{2}\lambda_0 \sum_{i=M_1(t)}^{M_2(t)}|\nu(\bar{r}^i)|(\bar{r}^i)^{n-1} - (n-1)\sum_{i=M_1(t)}^{M_2(t)}\nu(\bar{r}^i)(\bar{r}^i)^{n-2}.$$

We now want to pass to the limit $R_0 \to 0$. This is possible thanks to the following bounds. Since the jumps of v are exactly given by the interfaces R^i, with $\nu(R^i) \neq 0$, we have the formula

$$\sup_{(0,T)} \int_\Omega |\nabla v| = \sup_A \omega_n \sum_{i=1}^{\infty} |\nu(R^i)|(R^i)^{n-1} \leq C.$$

For the second sum, we apply Proposition 3.4. By lower semicontinuity, estimate (3.12) carries over to the limit $g(v) = \frac{4}{3}v$ and thus

$$\int_0^T \int_0^1 |v'| r^{n-2}\, dr\, dt = \int_A \sum_{i=1}^{\infty} |\nu(R^i)|(R^i)^{n-2}\, dt \leq C.$$

Thus we know that for almost any $t \in A$, all of the sums converge absolutely as $R_0 \to 0$, and by the monotone convergence theorem, this convergence (after multiplication with χ as in (6.23)) is in $L^1(0,T)$:

$$\left\{ \sum_{i=1}^{M_0} (-1)^{i+1} (\bar{r}^i)^{n-1} \dot{\bar{r}}^i \right\} \chi$$
$$\overset{R_0 \to 0}{\longrightarrow} \left\{ -\frac{3}{2}\lambda_0 \sum_{i=M_1(t)}^{\infty} |\nu(R^i)|(R^i)^{n-1} - (n-1) \sum_{i=M_1(t)}^{\infty} \nu(R^i)(R^i)^{n-2} \right\} \chi.$$

This together with (6.23) and (6.24) implies

$$\mu_{R_0} \overset{*}{\longrightarrow} \mu = 0$$

and

$$\frac{3}{2}\lambda_0 \sum_{i=M_1(t)}^{\infty} |\nu(R^i)|(R^i)^{n-1} + (n-1) \sum_{i=M_1(t)}^{\infty} \nu(R^i)(R^i)^{n-2} = 0$$

almost everywhere in A. □

We now summarize the results of this section in the following theorem.

THEOREM 6.6 (the limit equation). *Suppose that φ_ε is a smooth, radial solution of* (1.1) *with boundary condition* (1.2) *which satisfies* (2.1) *and* (2.2). *Then for the limit v of this sequence* (*cf. Remark* 2.4), *the free boundary $\Gamma := \partial\{v = -1\}$ is given by a collection of continuous graphs*

$$r^i : [t_i, T_i] \longrightarrow [0,1]$$

with the following conditions:

(1) $0 < r^i < 1$ *in* (t_i, T_i).

(2) r^i *is locally Lipschitz continuous in* $\{r^i > 0\} \cap [t_i, T_i]$.

(3) $\dot{r}^i \leq 0$ *almost everywhere in* (t_i, T_i).

(4) $t_i = 0$, $r^i(t_i) = 1$, *or* $t_i = t_j$ *for some* $j \neq i$ *and* $r^i(t_i) = r^j(t_j)$.

(5) $T_i = T$, $r^i(T_i) = 0$, *or* $T_i = T_j$ *for some* $j \neq i$ *and* $r^i(T_i) = r^j(T_j)$.

(6) *Two different graphs* r^i *and* r^j *agree at most at finitely many points.*

(7) *The direction of jump* $\nu^i(t) := \nu(t, r^i(t)) = \pm 1$ *is constant for* $t \in (t_i, T_i)$ *and—if the space dimension is larger than* 2 *or if the space dimension is* 2 *and the total number of interfaces is odd—then the multiplicity satisfies* $m_0(t, r^i(t)) = 1$.

(8) r^i *satisfies the nonlocal differential equation*

$$-\frac{4}{3}\left(\dot{r}^i + \frac{n-1}{r^i}\right) = 2\nu^i\lambda_0 \qquad in\ (t_i, T_i). \tag{6.25}$$

(9) *In* (8) *above, the Lagrange multiplier* λ_0 *is given by*

$$\lambda_0 = -(n-1)\frac{2}{3}\frac{\sum \nu^i (r^i)^{n-2}}{\sum (r^i)^{n-1}}$$

and changes sign exactly at times t_i *with* $r^i(t^i) = 1$.

The graphs r^i *correspond (up to renumbering) one to one to the true interfaces* R^j *given by* (6.22). *If the initial data have only finitely many true interfaces, than the number of true interfaces is finite at any given time. There might exist phantom interfaces which all evolve by mean curvature.*

Proof. By Proposition 6.3, we know that in $I^i_{R_0}$ the evolution equation is satisfied, and by Proposition 6.5, we have a formula for the Lagrange multiplier. Combining both propositions, we see

$$\dot{R}^i \le 0 \Longleftrightarrow \sum_j \nu(R^j)\nu(R^i)R^i(R^j)^{n-2} \le \sum_j |\nu(R^j)|(R^j)^{n-1}, \tag{6.26}$$

where summation is over all interfaces of radii less than 1. However, if k is the smallest index such that the corresponding interface has radius less than 1 and nonzero ν, then

$$\begin{aligned} \sum_j \nu(R^j)\nu(R^i)R^i(R^j)^{n-2} &\le \begin{cases} |\nu(R^i)|R^i(R^k)^{n-2} & \text{if } i \le k, \\ 0 \quad \text{otherwise} \end{cases} \\ &\le (R^k)^{n-1} \le \sum_j |\nu(R^j)|(R^j)^{n-1}. \end{aligned} \tag{6.27}$$

This proves that $\dot{R}^i \le 0$ in $I^i_{R_0}$ for any $R_0 > 0$.

Since mass is conserved, the denominator in the formula for λ_0 is bounded from below, and since the nominator is an alternating sum, λ_0 is uniformly bounded.

The sign of λ_0 is given by the sign of the first nonvanishing $\nu(R^j)$ that corresponds to some $R^j < 1$. Thus the sign of λ_0 changes only if some R^j with $\nu(R^j) \neq 0$ emerges from the fixed boundary. Since the number of interfaces larger than any R_0 is finite this may only occur at finitely many points.

Now suppose $t \in A_{R_0}$ for some R_0. Fix $R^j(t)$ and $R^k(t)$ with $j < k$ and assume that there exists for both $(R^j(t), t)$ and $(R^k(t), t)$ a neighborhood such that in these neighborhoods, $\nu(R^j) = \nu_j$ and $\nu(R^k) = \nu_k$ are constant and nonzero. Thus by the differential equation (6.18),

$$\begin{aligned} \dot{R}^j - \dot{R}^k &= \frac{n-1}{R^k} - \frac{n-1}{R^j} + \frac{3}{2}\,\lambda_0(\nu_k - \nu_j) \\ &> \frac{3}{2}\,\lambda_0(\nu_k - \nu_j), \end{aligned} \tag{6.28}$$

and hence the distance between R^j and R^k increases if $\lambda_0(\nu_k - \nu_j) \ge 0$. This is true in particular if $\nu_k = \nu_j$.

In the case where the dimension of the space $n \ge 3$, we know from (6.19) and Proposition 6.5 that for almost every $t \in A_{R_0}$, the condition $\nu(t, R^j(t)) \neq 0$ implies

$m_0 = 1$. By the definition of m_0, this implies that for almost all t and all points $(R^j(t), t)$ with $R^j(t) > R_0$ and $\nu(t, R^j(t)) \neq 0$, there exists a box B such that the set $\Gamma_{R_0} \cap B$ consists exactly of the corresponding R^j. Thus observation (6.28) implies that the distance between all interfaces R^k and R^j increases if they have the same normal, and hence such interfaces can never meet. As a consequence, we find that if two R^j and R^k with a nonvanishing normal meet at some point t, then $\nu_j - \nu_k$ is in either the set $\{2, 0\}$ or the set $\{0, -2\}$ in a whole neighborhood of the meeting point and consequently does not change sign. Thus (6.28) shows that either $\dot{R}^j - \dot{R}^k \geq 0$ or $\dot{R}^j - \dot{R}^k \leq 0$ in any one-sided neighborhood of the meeting point because the sign of λ_0 changes at most at finitely many times and $\frac{n-1}{R^k} - \frac{n-1}{R^j}$ is Lipschitz continuous in the neighborhood. Thus there are exactly three possibilities for geometric singularities in A_{R_0}: (i) two true interfaces meet and then form a single phantom interface; (ii) two interfaces nucleate out of a single phantom interface; or (iii) two true interfaces meet and immediately separate again. Phenomenon (iii) occurs only if λ_0 changes sign at that time point. Of course, phantom interfaces cannot meet each other because they move by mean curvature. Moreover, they can only meet any of the true interfaces where the latter cease to exist because otherwise the density m_0 would be larger than 1 on a true interface.

Similar arguments apply for $n = 2$. Indeed, in that case, the Lagrange multiplier locally either has a fixed sign or is identically zero. In the first case, the same analysis as above applies because the density m_0 of interfaces with nonvanishing normal is 1 again. In the second case, all the interfaces move by mean curvature and cannot meet anyway.

The limit v satisfies the weak Hölder-continuity estimate in Remark 2.11. For true interfaces, it implies that at the points of $N(R_0)$ (see (2.12) and Proposition 3.2 for the definition), only one of the following behaviors of the graphs R^j is possible.

At most two true interfaces can meet or nucleate at a point $t \in N(R_0)$. Two meeting true interfaces can only nucleate into two true interfaces across $t \in N(R_0)$ if λ_0 changes sign at this time point, and this can only happen if a true interface nucleates from the fixed boundary.

Any true interface that does not meet with another one at times in $N(R_0)$ continues as a single true interface across this time point.

When there exists a continuation, it has to be of class Hölder-$\frac{1}{2}$ and thus the differential equation is satisfied across that point.

These arguments allow us to conclude that the free boundary $\partial\{v = -1\}$ consists of a collection of graphs that have properties (1)–(10).

Thus the proof of Theorem 6.6 is complete. □

Remark 6.7. The evolution equation is the radial version of the expected limiting nonlocal geometric problem

$$V_i - k_i = -\frac{1}{\sum_j |\Gamma_j|} \sum_j \int_{\Gamma_j} k_j \, ds,$$

where V_i is the normal velocity and k_i is the sum of the principal curvatures of the interface Γ_i.

Note that in the right-hand side of (6.25), we sum only over interfaces R^i such that $\nu(R^i) \neq 0$. These interfaces represent exactly the free boundary of v. If $\nu(R^i) = 0$, these phantom interfaces are not seen by the limit v and they correspond to collapsing ε-interfaces. They do not have an impact on the evolution of the other interfaces.

Example 6.8. Colliding interfaces are generic to the nonlocal flow. Indeed, if two interfaces are sufficiently close to each other initially, sufficiently close with respect to their distance to the others and to their own size, then they have to meet before the smaller of them has time to shrink to 0.

Let $n = 3$ and assume that there are three initial interfaces. Then their evolution, as long as they all exist, is governed by

$$-\left(\dot{R}^i + \frac{n-1}{R^i}\right) = -(-1)^{i+1}(n-1)\mu_0,$$

where

$$\mu_0 = \frac{R^1 - R^2 + R^3}{(R^1)^2 + (R^2)^2 + (R^3)^2}.$$

Therefore

$$\dot{R}^3 \geq -\frac{n-1}{R^3}$$

and thus $(R^3)^2(t) \geq -2(n-1)t + (R^3)^2(0)$. Thus R^3 cannot disappear at the origin as long as $t \leq t_{\max} = \frac{1}{2(n-1)}(R^3)^2(0)$.

Since mass is preserved, the largest interface is bounded below by some number $R_{\min}$, which is such that the mass of a ball of radius $R_{\min}$ equals the initial mass:

$$(R_{\min})^3 = \left((R^1)^3 - (R^2)^3 + (R^3)^3\right)(0).$$

Since all interfaces are decreasing, this implies the following bound for μ_0:

$$\mu_0 \geq \frac{R_{\min} - R^2(0)}{((R^1)^2 + (R^2)^2 + (R^3)^2)(0)}.$$

Subtracting the equation for the second and third interface results in

$$\dot{R}^2 - \dot{R}^3 = \frac{n-1}{R^3} - \frac{n-1}{R^2} - 2(n-1)\mu_0 \leq -2(n-1)\mu_0.$$

Now suppose the interfaces R^2 and R^3 do not meet before R^3 vanishes. Then we can integrate the above inequality, use the bound for μ_0, and evaluate the result at $t_{\max}$ to find

$$0 \leq (R^2 - R^3)(t_{\max}) \leq -\frac{R_{\min} - R^2(0)}{((R^1)^2 + (R^2)^2 + (R^3)^2)(0)} + (R^2 - R^3)^2(0).$$

If $(R^2 - R^3)^2(0)$ is small, then $R_{\min}$ is approximately $R^1(0)$ and it is clearly possible to choose initial data for R^1 that contradicts the above inequality. Thus R^2 and R^3 have to meet before the smaller one has time to disappear. After the meeting point, the evolution becomes stationary and $R^1 = R_{\min}$.

Remark 6.9. Our estimates of section 5 are strong enough to prove that

$$E_0(t) = \frac{4}{3}\sum (R^i(t))^{n-1},$$

which contains contributions from all interfaces, both true and phantom. Thus jumps in the energy (which are responsible for the set of bad time points in the complement of A) correspond to either jumps in interfaces or jumps in multiplicity, the latter only possible for phantom interfaces. If we impose the condition that initially only finitely many and only true interfaces exist, then by the maximum principle for the ε-equation, no interfaces can nucleate from the origin as long as the smallest of the initial interfaces has not yet disappeared. Thus up to that time, only true interfaces exist and consequently have density 1 and are continuous.

Since the energy $E_*[v]$ of the limit (see (2.10) for the definition) counts only the true interfaces, we may identify any loss of energy in the limit as the appearance of phantom interfaces.

7. The viscous Cahn–Hilliard equation. Here we consider the viscous Cahn–Hilliard equation in $\Omega_T := \Omega \times (0,T)$ as introduced by Novick-Cohen [NC]:

$$\alpha\partial_t\varphi \;-\; \triangle u \;=\; 0, \tag{7.1}$$

$$u \;=\; -\varepsilon\triangle\varphi \;+\; \frac{1}{\varepsilon}W'(\varphi) \;+\; \nu\partial_t\varphi. \tag{7.2}$$

Imposing Neumann-zero boundary conditions for both u and φ and applying the usual techniques, one obtains the equations

$$\frac{1}{\alpha}\int_\Omega |\nabla u|^2\,dx \;+\; \partial_t E_\varepsilon(\varphi) \;+\; \nu\int_\Omega |\partial_t\varphi|^2\,dx \;=\; 0 \tag{7.3}$$

and

$$\int_\Omega \partial_t\varphi\,dx \;=\; 0 \quad\text{ and }\quad \fint_\Omega u\,dx \;=\; \frac{1}{\varepsilon}\fint_\Omega W'(\varphi)\,dx. \tag{7.4}$$

We shall show briefly how the viscous Cahn–Hilliard equation (7.1)–(7.2) relates both to the Cahn–Hilliard and the nonlocal Allen–Cahn equation.

Let us first consider the limit $\nu \to 0$, keeping all of the other parameters fixed. Formally, the limit problem is the standard Cahn–Hilliard equation. If the initial energy is uniformly bounded in ν, this can be shown rigorously. First, we note that estimate (7.3) immediately gives weak compactness in $L^2(\Omega_T)$ for $\nabla\varphi$ and a bound of φ in $L^4(\Omega_T)$. However, since the equation is nonlinear, it is important to have strong compactness in $L^1(\Omega_T)$ for φ. To this end, we note that equation (7.1) implies

$$\alpha||\partial_t\varphi||_{L^2(H^{-1,2})} \;=\; ||\nabla u||_{L^2(\Omega_T)} \;\le\; E_\varepsilon(\varphi)(0) \tag{7.5}$$

and thus (with $E_\varepsilon(\varphi)(0) \le C$)

$$\alpha||\varphi(\cdot,\cdot-h) - \varphi||_{L^2(H^{-1,2})} \;\le\; C\cdot h. \tag{7.6}$$

Now interpolating between $L^2(H^{-1,2})$ and $L^2(H^{1,2})$ yields

$$||\varphi(\cdot,\cdot+h) \;-\varphi||_{L^1(\Omega_T)} \;\le\; Ch^{\frac{1}{2}}. \tag{7.7}$$

This implies strong compactness for φ in $L^1(\Omega_T)$.

Of course the estimate for ∇u in (7.3) together with the bound of its mean value in (7.4) imply weak compactness in $L^2(\Omega_T)$ for both u and ∇u. This convergence is strong enough to pass to the limit $\nu \to 0$ in the viscous Cahn–Hilliard equation (7.1)–(7.2).

If the limit $\alpha \to 0$ is considered, the limit problem will be the nonlocal equation. To see this, assume once again that the energy is uniformly bounded in α. Then estimate (7.3) yields the L^1 compactness of the order parameter φ immediately, and the weak $L^2(\Omega_T)$ compactness of u and ∇u is again obtained from (7.3) and the mean-value condition (7.4). This compactness allows us to pass to the limit in (7.1) and (7.2). The limit of (7.1) gives $\triangle u = 0$ for the limit, and thus u is a constant, but then (7.4) yields the correct formula for u and we find the nonlocal equation in the limit.

Thus we see that both the nonlocal equation and the Cahn–Hilliard equation occur as special degenerate limits of the viscous Cahn–Hilliard equation.

Appendix. Ellipticity of the linearized Allen–Cahn equation. Here we give the proof of the ellipticity proposition (Proposition 5.6).

Proof of Proposition 5.6. Let S be one of the sets of integration as in Proposition 5.6. We start by integrating by parts the left-hand side of the claimed estimate (denoted (LHS)). This results in

$$\begin{aligned}\text{(LHS)} &= \int_S (-\Psi'' \ + \ W''(\Theta)\Psi)\Psi\xi^2 \, dz \\ &= \int_S \left(|\Psi'|^2 \ + \ W''(\Theta)\Psi^2\right)\xi^2 \, dz \ + \ 2\int_S \Psi'\Psi\xi'\xi \, dz. \end{aligned} \tag{A.1}$$

Here we should point out that Ψ is not necessarily smooth at $z = z_\varepsilon^i$, but since Ψ vanishes at these points, the integration by parts is nevertheless valid. Now we note that $W''(\Theta) = 2(3\Theta^2 - 1)$, and in the set where this function is strictly positive, there is nothing to prove. A careful study of Θ will show that the measure of the set where W'' is not strictly positive is small enough that the integral of Ψ^2 over this set can still be controlled by the integral of Ψ'^2.

Thus we introduce for any $a > 0$ the set

$$I_\varepsilon \ := \ \{W''(\Theta) \ < \ 2a\} \cap (-\varepsilon^{-\beta} + z_\varepsilon^{m_\varepsilon}, \varepsilon^{-\beta}). \tag{A.2}$$

We want to estimate the diameter of any connected component of I_ε. By the definitions in (5.5) and (5.6)

$$\Theta(z) \ = \ \sum_i \Xi_i(z)\tanh((-1)^i(z - z_\varepsilon^i) + \mu_\varepsilon) \ + \ \varepsilon\Phi_1^\varepsilon(z)$$

with

$$||\Phi_1||_{L^\infty(-\varepsilon^{-\beta}+z_\varepsilon^{m_\varepsilon}, \varepsilon^{-\beta})} \ \leq \ C||\lambda_\varepsilon|| \ \leq \ \frac{C}{\sqrt{\varepsilon}} \quad \text{uniformly in time}$$

by Corollary 2.2 and Lemma 5.7.

Since $W''(\Theta) < 2a \;\Leftrightarrow\; |\Theta| < \sqrt{\frac{a+1}{3}}$, we thus find that

$$\begin{aligned} I_\varepsilon &\subset \left\{ \sum_i \Xi_i |\tanh((-1)^i(z - z_\varepsilon^i) + \mu_\varepsilon)| \;\le\; \sqrt{\frac{a+1}{3}} \;+\; C\varepsilon \right\} \\ &\subset \bigcup_i \left\{ \;|\tanh((-1)^i(z - z_\varepsilon^i) + \mu_\varepsilon)| \;\le\; \sqrt{\frac{a+1}{3}} \;+\; C\varepsilon \right\} \\ &\subset \bigcup_i \left\{ |z - z_\varepsilon^i| \;\le \tanh^{-1}\left(\sqrt{\frac{a+1}{3}} \;+ C\varepsilon \right) \;+\; |\mu_\varepsilon| \right\} \\ &=: \bigcup_i I^i. \end{aligned} \tag{A.3}$$

We return to (A.1). We continue to estimate as follows:

$$\begin{aligned} \text{(LHS)} \;\ge\; &\int_S |\Psi'|^2 \xi^2 \, dz \;+\; 2a \int_S \Psi^2 \xi^2 \, dz \\ &- (2+2a) \int_{I_\varepsilon} \Psi^2 \xi^2 \, dz \;+\; 2 \int_S \Psi' \Psi \xi' \xi \, dz. \end{aligned} \tag{A.4}$$

However, using $\Psi(z_\varepsilon^i) = 0$ gives

$$\begin{aligned} \int_{I_\varepsilon} \Psi^2 \xi^2 \, dz = \; &\sum_i \int_{I^i \setminus I^{i-1}} \xi^2 \Psi^2 \, dz \\ = \; &\sum_i \int_{I^i \setminus I^{i-1}} \xi^2 \left(\int_{z_\varepsilon^i}^z \Psi' \right)^2 \, dz \\ \le \; &\sum_i \int_{I^i \setminus I^{i-1}} |\Psi'|^2 \frac{1}{2} |I^i| \\ \le \; &\left(\tanh^{-1}\left(\sqrt{\frac{a+1}{3}} + C\varepsilon \right) \;+\; \tanh^{-1} Q \right) \int_S |\Psi'|^2 \xi^2 \, dz \end{aligned} \tag{A.5}$$

by (A.3) and since $|\mu_\varepsilon| = |\tanh^{-1} a_\varepsilon| \le \tanh^{-1} Q$. Note in addition that by construction $\bigcup I^i \subset \{\xi = 1\}$. We enter this into (A.4) and continue to estimate:

$$\begin{aligned} (LHS) \;\ge\; &\underbrace{\left(1 \;-\; (2+2a) \left(tanh^{-1} \left(\sqrt{\frac{a+1}{3}} + C\varepsilon \right) \;+\; tanh^{-1} Q \right) \right)}_{=: c_0(a,\varepsilon)} \int_S |\Psi'|^2 \xi^2 \, dz \\ &+\; 2a \int_S \Psi^2 \xi^2 \, dz \;+\; 2 \int_S \Psi' \Psi \xi' \xi \, dz. \end{aligned} \tag{A.6}$$

The number c_0 turns out to be positive if $\varepsilon = 0$ and $a = 0$ by the choice of Q in (4.1), and thus there exist positive ε_0 and a_0 such that

$$0 \;<\; c_0(\varepsilon_0, a_0) \;<\; c_0(\varepsilon, a_0)$$

for all $\varepsilon < \varepsilon_0$.

This proves the proposition with $\zeta_1 = \frac{1}{2} c_0(\varepsilon_0, a_0)$ and $\zeta_2 = 2a_0$ if we still use the Hölder estimate for $\int \Psi' \Psi \xi' \xi$. □

Acknowledgment. The first author would like to thank the SFB256 in Bonn for the very warm and enriching environment they have created for their visitors.

REFERENCES

[ABC] N. Alikakos, P. Bates, and X. Chen, *Convergence of the Cahn–Hilliard equation to the Hele–Shaw model*, Arch. Rational Mech. Anal., 128 (1994), pp. 165–205.

[ABF] N. Alikakos, P. Bates, and G. Fusco, *Slow motion for the Cahn–Hilliard equation in one space dimension*, J. Differential Equations, 90 (1991), pp. 81–135.

[AC] S. Allen and J. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica, 27 (1979), pp. 1084–1095.

[AW] D. Aronson and H. Weinberger, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, Springer-Verlag, New York, 1975, pp. 5–49.

[BF] M. Berger and L. Fraenkel, *On the asymptotic solution of a nonlinear Dirichlet problem*, J. Math. Mech., 19 (1970), pp. 553–585.

[BH] L. Bronsard and D. Hilhorst, *On the slow dynamics for the Cahn–Hilliard equation in one space dimension*, Proc. Roy. Soc. London Sect. A, 439 (1992), pp. 669–682.

[BK1] L. Bronsard and R. V. Kohn, *On the slowness of phase boundary motion in one space dimension*, Comm. Pure Appl. Math., 43 (1990), pp. 983–997.

[BK2] L. Bronsard and R. V. Kohn, *Motion by mean curvature as the singular limit of Ginzburg–Landau dynamics*, J. Differential Equations, 90 (1991), pp. 211–237.

[BX] P. Bates and J. Xun, *Metastable patterns for the Cahn–Hilliard equation* I, J. Differential Equations, 111 (1994), pp. 421–457.

[BS] L. Bronsard and B. Stoth, *On the existence of high multiplicity interfaces*, Math. Res. Lett., 3 (1996) pp. 41–50.

[CH] J. Cahn and J. Hilliard, *Free energy of a non-uniform system* I: *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[C] X. Chen, *Hele–Shaw problem and area–preserving, curve shortening motion*, Arch. Rational Mech. Anal., 123 (1993), pp. 117–151.

[CGG] Y.-G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 73 (1991), pp. 749–786.

[DS1] P. DeMottoni and M. Schatzman, *Evolution géometrique d'interface,* C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 453–458.

[DS2] P. DeMottoni and M. Schatzman, *Geometrical evolution of developed interfaces*, Trans. Amer. Math. Soc., (1995), pp. 1533–1589.

[ESS] L. Evans, M. Soner, and P. Souganidis, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., 45 (1992), pp. 1097–1123.

[ES] L. Evans and J. Spruck, *Motion of level sets by mean curvature* I, J. Differential Geom., 37 (1991), pp. 635–681.

[F] P. Fife, *Models for phase separation and their mathematics*, in Proc. Taniguchi International Symposium on Nonlinear Partial Differential Equations and Applications, Kinokuniya, 1991.

[FM] P. Fife and B. McLeod, *The approach of solutions of nonlinear diffusion equation to traveling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[Ga] M. Gage, *On an area–preserving evolution for planar curves*, Contemp. Math., 51 (1986), pp. 51–62.

[G] C. Grant, *Slow motion in one dimensional Cahn–Morrel systems*, SIAM J. Math. Anal., 26 (1995), pp. 21–34.

[H] G. Huisken, *The volume preserving mean curvature flow*, J. Reine Angew. Math., 382 (1987), pp. 35-48.

[LM] S. Luckhaus and L. Modica, *The Gibbs–Thomson relation within the gradient theory for phase transitions*, Arch. Rat. Mech. Anal., 107 (1989), pp. 71–83.

[M] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[MM] L. Modica and S. Mortola, *Il limite nella* Γ *convergenza di una famiglia di funzionali elliptichi*, Boll. Un. Mat. Ital. A, 14 (1977), pp. 526–529 (in Italian).

[MS] W. Mullins and R. Sekerka, *Morphological stability of a particle growing by diffusion and heat flow*, J. Appl. Phys., 34 (1963), pp. 323–329.

[N] B. Niethammer, *Existence and uniqueness of radially symmetric stationary points within* *the gradient theory of phase transitions*, European J. Appl. Math., 6 (1995), pp. 45–67.

[NC] A. Novick-Cohen, *On the viscous Cahn–Hilliard equation*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J. Ball, ed., Oxford Scientific, Oxford, UK, 1988, pp. 329–342.

[P] R. Pego, *Front migration in the nonlinear Cahn–Hilliard equation*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 261–278.

[RS] J. Rubinstein and P. Sternberg, *Nonlocal reaction diffusion equations and nucleation*, IMA J. Appl. Math., 48 (1992), pp. 249–264.

[So] H. M. Soner, *Ginzburg–Landau equation and motion by mean curvature* I: *Convergence,* II: *Development of the initial interface*, J. Geom. Anal., to appear.

[S] P. Sternberg, *The effect of a singular pertubation on nonconvex variational problems*, Arch. Rational Mech. Anal., 101 (1988), pp. 209–260.

[S1] B. Stoth, *A sharp interface limit of the phase-field equations: One–dimensional and axisymmetric*, European J. Appl. Math., 7 (1996), pp. 603–633.

[S2] B. Stoth, *Convergence of the Cahn–Hilliard equation to the Mullins–Sekerka problem in spherical symmetry*, J. Differential Equations, 125 (1996), pp. 154–183.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 808–830, July 1997

SPATIAL DECAY ESTIMATES FOR FLOW IN A POROUS MEDIUM*

J. CHADAM† AND Y. QIN‡

Abstract. In the spirit of the recent work of Ames, Payne, and Schaefer [*SIAM J. Math. Anal.*, 24 (1993), pp. 97–116], the decay of flow in a porous medium through a semi-infinite pipe is investigated. The analysis presented in this paper is based on the Brinkman–Forchheimer model. In establishing decay estimate and bounds for the total weighted energy, the nonlinear term in the model equation leads to difficulties which were not encountered in the paper cited above.

Key words. Brinkman–Forchheimer flow, decay estimates

AMS subject classifications. 35B45, 35Q30, 76D07

PII. S0036141095290562

1. Introduction.

In a recent paper [1], Ames et al. investigated the time-dependent Stokes flow of an incompressible viscous fluid in a semiinfinite cylindrical pipe. An exponential decay of a weighted energy expression was derived if the net flow through the finite end of the pipe is assumed to be zero for each t. Modeled on their work, we consider an analogous problem for flow in a porous medium based on a time-dependent nonlinear equation, namely the Brinkman–Forchheimer equation. While non-Darcy models in porous media have been increasingly discussed in the recent engineering literature, the related mathematical analysis for some problems based on these models is much less developed. In this paper, we shall study the spatial decay estimates for the Brinkman–Forchheimer flow.

Although the techniques used in the present study basically follow the suggestions in [1], the work reported here is not a trivial extension. Since the Stokes equation is a linear equation, the interesting feature of this paper is the extension of the analysis of decay estimates to a time-dependent nonlinear problem of pipe flow. An important point of our analysis is to ensure that the estimates depend only on the known data, i.e., the physical parameters, the initial and boundary data, and the geometry of the domain. The presence of the nonlinear term makes the problem considerably more complicated and new techniques must be developed to handle these difficulties. We obtain the bound by two steps. First, we compare the energy of the solution to the Brinkman–Forchheimer problem with the energy of the solution to the linearized Brinkman problem. We then show how to determine bounds on the energy of the linearized problem and the total energy bound to the nonlinear problem in a manner which depends only on the data. Our motivation comes from the earlier work of Horgan and Wheeler [10] and Ames and Payne [2] in studying decay estimates of steady pipe flow. The basic idea is originally due to Payne [14] in his investigation of the uniqueness criteria for steady-state solutions of the Navier–Stokes equation. Spatial decay estimates for similar flow in a planar region have been studied in [16]. This analysis depended heavily on the use of a stream function which could be obtained

[*] Received by the editors August 21, 1995; accepted for publication (in revised form) February 28, 1996. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sima/28-4/29056.html

[†] The Fields Institute, 222 College Street, Toronto, ON M5T 3J1, Canada. Current address: Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (chadam+@pitt.edu).

[‡] Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Toronto, ON L8S 4K1, Canada (qinyu@mcmail.mcmaster.ca).

through the elimination of the pressure. This trick is not possible in the present case of pipe flow and the interest here is to extend the analysis of the decay estimates without recourse to these methods.

For investigations of the spatial decay of solutions of elliptic and parabolic equations, see, e.g., [3, 4, 5, 6, 7, 9, 12] and the references cited therein.

In the next section, we formulate the initial-boundary value problem, which describes pipe flow in a porous medium. We derive a second-order differential inequality for the energy expression in section 3. A method for finding bounds on the estimates for both the nonlinear problem and corresponding linear problem is given in section 4. For comparison, we discuss the decay results for Darcy pipe flow in the last section.

2. Formulation of the problem. We begin this section by first introducing the model to be used in this paper. Most of the studies in porous media carried out to date have been based on the Darcy law, which is an empirical law for creeping flow through an infinitely extended uniform medium. However, researchers now generally recognize that non-Darcian effects are quite important for certain applications. For a discussion of generalized non-Darcy models, see Vafai and Tien [17], Hsu and Chen [11], and Prasad and Kladias [15]. Despite a difference of opinions as to what is the most appropriate form for modeling flow in porous media, the Brinkman–Forchheimer equation has wide acceptance in the porous media literature. The time-dependent momentum equation describing flow in a porous medium may be written as [13]

$$\rho_f c_a \frac{\partial \mathbf{u}}{\partial t} = -\nabla p_f + \mu_{\text{eff}} \nabla^2 \mathbf{u} - \frac{\mu_f}{k}\mathbf{u} - \frac{F\rho_f \phi}{k^{\frac{1}{2}}}|\mathbf{u}|\mathbf{u}, \tag{2.1}$$

where ρ_f is the density of the fluid, c_a is the acceleration coefficient, p_f is the intrinsic fluid pressure, $\mathbf{u}$ is the Darcy velocity, μ_f is the viscosity, μ_{eff} is the effective viscosity, k is the permeability, F is the Forchheimer coefficient, ϕ is the porosity, and $|\mathbf{u}| = \sqrt{u_1^2 + u_2^2 + u_3^2}$.

After scaling, the nondimensional form of (2.1) is

$$\mathbf{u}_t = \Delta \mathbf{u} - J_1 \mathbf{u} - J_2 |\mathbf{u}|\mathbf{u} - \nabla p, \tag{2.2}$$

where Δ is the three-dimensional Laplacian and J_1 and J_2 are two physical parameters, the first being the coefficient of the Darcy velocity and the second being the coefficient of inertial drag for flow in a porous medium.

Let R denote the interior of a semiinfinite cylindrical pipe of an arbitrary, smooth cross-section with generator parallel to the x_3-axis. The end (entrance) of the pipe in the $x_3 = 0$ plane is denoted by D_0 and comprises part of the boundary ∂R of R.

We let

$$R_z = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D_0, x_3 > z \geq 0\}$$

denote the subdomain of R for which $x_3 > z$ and let

$$D_z = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D_0, x_3 = z\}$$

denote the part of ∂R_z in the plane $x_3 = z \geq 0$.

The initial-boundary value problem that we consider here is

$$u_{i,t} = \Delta u_i - J_1 u_i - J_2 |\mathbf{u}| u_i - p_{,i} \quad \text{in } R \times (0, \infty), \tag{2.3}$$

$$u_{i,i} = 0 \quad \text{in} \quad \bar{R} \times (0,\infty), \tag{2.4}$$

$$u_i = 0 \quad \text{on} \quad \partial R \backslash D_0 \times (0,\infty), \tag{2.5}$$

$$u_i = f_i(x_1,\, x_2,\, t) \quad \text{in} \quad \bar{D}_0 \times (0,\infty), \tag{2.6}$$

$$u_i = 0 \quad \text{in } R \times \{0\}, \tag{2.7}$$

$$u_i,\, u_{i,j},\, u_{i,t},\, p = o(x_3^{-1}) \quad \text{uniformly in } x_1,\, x_2,\, t \quad \text{as } x_3 \to \infty. \tag{2.8}$$

Here the comma (partial differentiation) and repeated index (summation) conventions are used. In this work, Latin subscripts range over 1, 2, 3 while Greek subscripts range over 1, 2 unless otherwise noted. We assume that the velocity field $u_i(x_1,\, x_2,\, x_3,\, t)$ for $i = 1,\, 2,\, 3$ and the pressure $p(x_1,\, x_2,\, x_3,\, t)$ are classical solutions of (2.3)–(2.8) and the prescribed functions (entrance profile) f_i are continuously differentiable and vanish on $\partial D_0 \times [0,\, \infty)$. The notation $|\mathbf{u}|$ represents $(u_i u_i)^{\frac{1}{2}}$.

We note that in the special case when $J_1 = 0$ and $J_2 = 0$ (flow in a clear fluid), the problem (2.3)–(2.8) is reduced to that considered in [1].

We define a weighted energy integral for the solution u_i of (2.3)–(2.8) by (no summation on τ)

$$E(z,\, t) = \int_0^t \int_{R_z} (\xi - z)[J_1 u_i u_i + u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}]\, dx d\tau, \tag{2.9}$$

with k a parameter to be chosen later, and note that

$$\frac{\partial E}{\partial z} = -\int_0^t \int_{R_z} [J_1 u_i u_i + u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}]\, dx d\tau, \tag{2.10}$$

$$\frac{\partial^2 E}{\partial z^2} = \int_0^t \int_{D_z} [J_1 u_i u_i + u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}]\, dA d\tau. \tag{2.11}$$

Our task is to obtain a second-order differential inequality for E from which we can deduce an exponential decay estimate.

We remark here that from the point of view of the mathematics, it is not necessary to include the lower-order term of the velocity (the first term in the integrand of integral (2.9)) in the weighted energy because the integral of this term can be controlled by the integral of the spatial derivative term using the Poincaré inequality (see (A.1) in the appendix). However, the parameter J_1 is a relatively large number for most porous media occurring in nature (whereas it is small for a sparse material with many pores), particularly in the limiting case of Darcy's law. Thus the square of the L^2 norm of u_i is the only kinetic energy term (see section 5) and we would like to retain this term in our weighted energy.

3. Energy estimation. We now begin to derive the desired differential inequality for E from which the decay results will follow.

On multiplying (2.3) by $(\xi - z)u_i$ and $k(\xi - z)u_{i,t}$, integrating over R_z, and using the boundary conditions for u_i, it can be shown that

$$\begin{aligned} E(z,t) = &-\int_0^t \int_{R_z} u_i u_{i,3}\, dx d\tau - \int_0^t \int_{R_z} (\xi - z) u_i [u_{i,\tau} + J_2 |\mathbf{u}| u_i + p_{,i}]\, dx d\tau \\ &+ k \int_0^t \int_{R_z} (\xi - z) u_{i,\tau} [\Delta u_i - J_1 u_i - J_2 |\mathbf{u}| u_i - p_{,i}]\, dx d\tau. \end{aligned}$$

Integrating by parts again, using (2.4) and the initial-boundary condition, and dropping five terms which are negative, we have

$$\begin{aligned} E(z,t) &\leq -\int_0^t \int_{R_z} u_i u_{i,3}\, dx d\tau - k\int_0^t \int_{R_z} u_{i,\tau} u_{i,3}\, dx d\tau \\ &\quad + \int_0^t \int_{R_z} u_3 p\, dx d\tau + k\int_0^t \int_{R_z} u_{3,\tau} p\, dx d\tau \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{3.1}$$

Estimates of the first two integral terms can easily be obtained by use of the Schwarz, Poincaré, and arithmetic mean–geometric mean (AG) inequalities

$$I_1 \leq \frac{1}{2\sqrt{\lambda_1}} \int_0^t \int_{R_z} u_{i,j} u_{i,j}\, dx d\tau, \tag{3.2}$$

$$I_2 \leq \frac{1}{2\sqrt{\lambda_1}} \left\{ \int_0^t \int_{R_z} k u_{i,\tau} u_{i,\tau}\, dx d\tau + \int_0^t \int_{R_z} u_{i,3} u_{i,3}\, dx d\tau \right\}. \tag{3.3}$$

In (3.3), we have selected $k = 1/\lambda_1$. Thus we have shown that

$$I_1,\ I_2 \leq \frac{1}{2\sqrt{\lambda_1}} \left(-\frac{\partial E}{\partial z} \right). \tag{3.4}$$

We now bound I_3 using the method of [1]. We first note that for any $z^* > 0$, by (2.4) and (2.5),

$$\begin{aligned} \int_{D_z} u_3\, dA &= \int_{D_z^*} u_3\, dA - \int_z^{z^*} \int_{D_\xi} u_{3,3}\, dA d\xi \\ &= \int_{D_{z^*}} u_3\, dA + \int_z^{z^*} \int_{D_\xi} u_{3,3}\, dA d\xi = \int_{D_{z^*}} u_3\, dA, \end{aligned}$$

and thus the area mean value of u_3 is the same over each section. Since u_3 is assumed to vanish at infinity, we conclude that this value is zero for all $z \geq 0$ and hence requires that f_3 satisfy

$$\int_{D_0} f_3\, dA = 0 \quad \text{for all } t \geq 0. \tag{3.5}$$

Under this assumption, there exists a vector function (see (A.3) in the appendix) ω_α which satisfies

$$\begin{aligned} I_3 &= \int_0^t \int_{R_z} \omega_{\alpha,\alpha} p\, dx d\tau = -\int_0^t \int_{R_z} \omega_\alpha p_{,\alpha}\, dx d\tau \\ &= \int_0^t \int_{R_z} \omega_\alpha [u_{\alpha,\tau} - \Delta u_\alpha + J_1 u_\alpha + J_2 |\mathbf{u}| u_\alpha]\, dx d\tau \\ &= \int_0^t \int_{R_z} \omega_\alpha u_{\alpha,\tau}\, dx d\tau + \int_0^t \int_{R_z} \omega_{\alpha,i} u_{\alpha,i}\, dx d\tau + \int_0^t \int_{D_z} \omega_\alpha u_{\alpha,3}\, dA d\tau \\ &\quad + J_1 \int_0^t \int_{R_z} \omega_\alpha u_\alpha\, dx d\tau + J_2 \int_0^t \int_{R_z} \omega_\alpha |\mathbf{u}| u_\alpha\, dx d\tau \\ &= I_{31} + I_{32} + I_{33} + I_{34} + I_{35}. \end{aligned} \tag{3.6}$$

The integrals of the auxiliary function ω_α can be bounded in terms of u_3 by means of (A.1) and (A.3). We can write

$$\begin{aligned}\int_0^t\int_{R_z}\omega_\alpha\omega_\alpha\,dxd\tau &\le \frac{1}{\lambda_1}\int_0^t\int_{R_z}\omega_{\alpha,\beta}\omega_{\alpha,\beta}\,dxd\tau\\ &\le \frac{C}{\lambda_1}\int_0^t\int_{R_z}(u_3)^2\,dxd\tau, \qquad (3.7)\\ \int_0^t\int_{R_z}\omega_{\alpha,3}\omega_{\alpha,3}\,dxd\tau &\le \frac{1}{\lambda_1}\int_0^t\int_{R_z}\omega_{\alpha,\beta 3}\omega_{\alpha,\beta 3}\,dxd\tau\\ &\le \frac{C}{\lambda_1}\int_0^t\int_{R_z}(u_{3,3})^2\,dxd\tau. \qquad (3.8)\end{aligned}$$

Using the Schwarz, Poincaré, and weighted AG inequalities, (3.7) and (3.8), we have the following estimates for the first three integrals (see [1]):

$$\begin{aligned}I_{31}+I_{32}+I_{33} &\le \frac{1}{2}\sqrt{\frac{C}{\lambda_1}}\left\{k\int_0^t\int_{R_z}u_{\alpha,\tau}u_{\alpha,\tau}\,dxd\tau+\int_0^t\int_{R_z}u_{3,\beta}u_{3,\beta}\,dxd\tau\right\}\\ &\quad+\frac{1}{2}\sqrt{\frac{C}{\lambda_1}}\int_0^t\int_{R_z}u_{i,j}u_{i,j}\,dxd\tau\\ &\quad+\frac{\sqrt{C}}{2\lambda_1}\left\{\int_0^t\int_{D_z}u_{\alpha,3}u_{\alpha,3}\,dxd\tau+\int_0^t\int_{D_z}u_{3,\beta}u_{3,\beta}\,dxd\tau\right\}. \qquad (3.9)\end{aligned}$$

For the fourth term on the right side of (3.6), we have

$$\begin{aligned}I_{34} &= J_1\int_0^t\int_{R_z}\omega_\alpha u_\alpha\,dxd\tau\\ &\le J_1\Big(\int_0^t\int_{R_z}\omega_\alpha\omega_\alpha\,dxd\tau\Big)^{\frac{1}{2}}\Big(\int_0^t\int_{R_z}u_\alpha u_\alpha\,dxd\tau\Big)^{\frac{1}{2}}\\ &\le J_1\sqrt{\frac{C}{\lambda_1}}\Big(\int_0^t\int_{R_z}(u_3)^2\,dxd\tau\Big)^{\frac{1}{2}}\Big(\int_0^t\int_{R_z}u_\alpha u_\alpha\,dxd\tau\Big)^{\frac{1}{2}}\\ &\le \frac{J_1}{2}\sqrt{\frac{C}{\lambda_1}}\int_0^t\int_{R_z}u_iu_i\,dxd\tau. \qquad (3.10)\end{aligned}$$

Now we need to estimate the last integral, I_{35}:

$$\begin{aligned}I_{35} &= J_2\int_0^t\int_{R_z}w_\alpha|\mathbf{u}|u_\alpha\,dxd\tau\\ &\le J_2\Big(\int_0^t\int_{R_z}\omega_\alpha\omega_\alpha\,dxd\tau\Big)^{\frac{1}{2}}\Big(\int_0^t\int_{R_z}(u_iu_i)^2\,dxd\tau\Big)^{\frac{1}{2}}. \qquad (3.11)\end{aligned}$$

By means of the Hölder inequality and the Sobolev inequality ((A.5) in the appendix) with $w = (u_iu_i)^{\frac{1}{2}}$ and $\bar{\Omega} = 4\Omega$, we deduce that

$$\begin{aligned}
&\int_0^t \int_{R_z} (u_iu_i)^2 \, dx d\tau \\
&\quad\leq \int_0^t \left(\int_{R_z} (u_iu_i)^{\frac{3}{2}} \, dx \right)^{\frac{2}{3}} \left(\int_{R_z} (u_iu_i)^3 \, dx \right)^{\frac{1}{3}} d\tau \\
&\quad\leq \max_{0\leq\tau\leq t} \left(\int_{R_z} (u_iu_i)^{\frac{3}{2}}(\tau) \, dx \right)^{\frac{2}{3}} \int_0^t \left(\int_{R_z} (u_iu_i)^3 \, dx \right)^{\frac{1}{3}} d\tau \\
&\quad\leq \Omega^{\frac{1}{3}} \max_{0\leq\tau\leq t} \left(\int_{R_z} (u_iu_i)^{\frac{3}{2}}(\tau) \, dx \right)^{\frac{2}{3}} \int_0^t \int_{R_z} u_{i,j}u_{i,j} \, dx d\tau. \qquad (3.12)
\end{aligned}$$

We point out here that to apply (A.5), we have extended u_i as an even function across $x_3 = z$. In fact, let $\tilde{w}$ be an even extension of w about $x_3 = z$, defined by

$$\tilde{w}(x_1, x_2, x_3, t) = \begin{cases} w(x_1, x_2, x_3, t) & \text{if } x_3 \geq z, \\ w(x_1, x_2, 2z - x_3, t) & \text{if } x_3 < z. \end{cases}$$

The Sobolev inequality ((A.5)) provides that

$$\int_{-\infty}^{\infty} \int_{D_\xi} \tilde{w}^6 \, dA d\xi \leq \bar{\Omega} \left(\int_{-\infty}^{\infty} \int_{D_\xi} \tilde{w}_{,j} \tilde{w}_{,j} \, dA d\xi \right)^3.$$

Using the substitution, we have

$$\int_{-\infty}^{\infty} \int_{D_\xi} \tilde{w}^6 \, dA d\xi = 2 \int_z^{\infty} \int_{D_\xi} w^6 \, dA d\xi = 2 \int_{R_z} w^6 \, dx.$$

$$\int_{-\infty}^{\infty} \int_{D_\xi} \tilde{w}_{,j} \tilde{w}_{,j} \, dA d\xi = 2 \int_{R_z} w_{,j} w_{,j} \, dx.$$

Replacing $w = (u_iu_i)^{\frac{1}{2}}$ and further using the Schwarz inequality, we thus deduce that

$$\int_{R_z} (u_iu_i)^3 \, dx \leq \Omega \left(\int_{R_z} u_{i,j}u_{i,j} dx \right)^3,$$

which justifies the last step of (3.2) in using the Sobolev inequality with $\bar{\Omega} = 4\Omega$. Substituting the estimate of (3.12) into (3.11) and employing (3.7) and (A.1), we conclude that

$$I_{35} \leq \frac{J_2 \Omega^{\frac{1}{6}} C^{\frac{1}{2}}}{\lambda_1} \max_{0\leq\tau\leq t} \left(\int_{R_z} (u_iu_i)^{\frac{3}{2}}(\tau) \, dx \right)^{\frac{1}{3}} \int_0^t \int_{R_z} u_{i,j}u_{i,j} \, dx d\tau. \qquad (3.13)$$

Finally, summarizing all of the results in (3.9), (3.10), and (3.13), we obtain

$$I_3 \leq \left\{ \sqrt{\frac{C}{\lambda_1}} + \frac{J_2 \Omega^{\frac{1}{6}} C^{\frac{1}{2}}}{\lambda_1} \max_{0\leq\tau\leq t} \left(\int_{R_z} (u_iu_i)^{\frac{3}{2}}(\tau) \, dx \right)^{\frac{1}{3}} \right\} \left(-\frac{\partial E}{\partial z} \right) + \frac{\sqrt{C}}{2\lambda_1} \frac{\partial^2 E}{\partial z^2}. \qquad (3.14)$$

We now turn to bounding I_4 in terms of $\partial E/\partial z$ and $\partial^2 E/\partial z^2$. Following the ideas in [1], we accomplish this with the aid of two auxiliary functions.

Let φ be a solution of the boundary value problem

$$\Delta\varphi = u_{3,t} \quad \text{in } R_z, \qquad \frac{\partial\varphi}{\partial\mathbf{n}} = 0 \quad \text{on } \partial R_z. \tag{3.15}$$

Since

$$\int_{R_z} u_{3,t}\,dx = \int_z^\infty \int_{D_\xi} u_{3,t} dA d\xi = 0$$

and $u_{3,t} \to 0$ as $x_3 \to \infty$, there exists a solution φ which together with its spatial derivatives vanishes as $x_3 \to \infty$. By means of φ, we can write

$$\begin{aligned} I_4 &= k\int_0^t \int_{R_z} p\Delta\varphi\,dxd\tau = -k\int_0^t \int_{R_z} \varphi_{,i}p_{,i}\,dxd\tau \\ &= k\int_0^t \int_{R_z} \varphi_{,i}\big[u_{i,\tau} - (u_{i,j} - u_{j,i})_{,j} + J_1 u_i + J_2|\mathbf{u}|u_i\big]\,dxd\tau \\ &= k\int_0^t \int_{R_z} \varphi_{,i}u_{i,\tau}\,dxd\tau + k\int_0^t \int_{D_z} \varphi_{,\alpha}(u_{\alpha,3} - u_{3,\alpha})\,dAd\tau \\ &\quad - k\int_0^t \int_z^\infty \int_{\partial D_\xi} \varphi_{,i}(u_{i,j} - u_{j,i})n_j\,dsd\xi d\tau \\ &\quad + kJ_1\int_0^t \int_{R_z} \varphi_{,i}u_i\,dxd\tau + kJ_2\int_0^t \int_{R_z} \varphi_{,i}|\mathbf{u}|u_i\,dxd\tau, \end{aligned}$$

and by Schwarz's inequality, we have

(3.16)

$$\begin{aligned} I_4 \le & k\Big(\int_0^t \int_{R_z} \varphi_{,i}\varphi_{,i}\,dxd\tau\Big)^{\frac12}\Big(\int_0^t \int_{R_z} u_{i,\tau}u_{i,\tau}\,dxd\tau\Big)^{\frac12} \\ &+ k\Big(\int_0^t \int_{D_z} \varphi_{,\alpha}\varphi_{,\alpha}\,dAd\tau\Big)^{\frac12}\Big(\int_0^t \int_{D_z} (u_{\alpha,3} - u_{3,\alpha})(u_{\alpha,3} - u_{3,\alpha})\,dAd\tau\Big)^{\frac12} \\ &+ k\Big(\int_0^t \int_z^\infty \int_{\partial D_\xi} |\mathrm{grad}_s\varphi|^2\,dsd\xi d\tau\Big)^{\frac12}\Big(\int_0^t \int_z^\infty \int_{\partial D_\xi} (u_{i,j} - u_{j,i})u_{i,j}\,dsd\xi d\tau\Big)^{\frac12} \\ &+ kJ_1\Big(\int_0^t \int_{R_z} \varphi_{,i}\varphi_{,i}\,dxd\tau\Big)^{\frac12}\Big(\int_0^t \int_{R_z} u_iu_i\,dxd\tau\Big)^{\frac12} \\ &+ kJ_2\Big(\int_0^t \int_{R_z} \varphi_{,i}\varphi_{,i}\,dxd\tau\Big)^{\frac12}\Big(\int_0^t \int_{R_z} (u_iu_i)^2\,dxd\tau\Big)^{\frac12}, \end{aligned}$$

where $\mathrm{grad}_s\varphi$ denotes the tangential component of the gradient of φ. The integrals involving the auxiliary function φ can be estimated by using (A.6)–(A.8) in the appendix with $f = u_{3,t}$. Using the weighted AG inequality and (3.12), we deduce that

$$\begin{aligned}
I_4 \leq & \left(\frac{3\sqrt{\lambda_1} + J_1^{\frac{1}{2}} + J_2}{2\sqrt{\lambda_1 \mu_2}} + \frac{2}{h_0^2 \sqrt{\lambda_1}} \left[\frac{1}{\mu_2} + \frac{d^2}{4} \right] \right) \int_0^t \int_{R_z} k u_{i,\tau} u_{i,\tau} \, dx d\tau \\
& + \frac{1}{\lambda_1} \int_0^t \int_{D_z} (u_{\alpha,3} - u_{3,\alpha})(u_{\alpha,3} - u_{3,\alpha}) \, dA d\tau \\
& + \frac{h_0}{4\sqrt{\lambda_1}} \int_0^t \int_z^\infty \int_{D_\xi} (u_{i,j} - u_{j,i}) u_{i,j} \, ds d\xi d\tau + \frac{J_1^{\frac{1}{2}}}{2\sqrt{\lambda_1 \mu_2}} \int_0^t \int_{R_z} J_1 u_i u_i \, dx d\tau \\
& + \frac{J_2 \Omega^{\frac{1}{3}}}{2\sqrt{\lambda_1 \mu_2}} \max_{0 \leq \tau \leq t} \left(\int_{R_z} (u_i u_i)^{\frac{3}{2}} (\tau) \, dx \right)^{\frac{2}{3}} \int_0^t \int_{R_z} u_{i,j} u_{i,j} \, dx d\tau.
\end{aligned} \tag{3.17}$$

In order to estimate the third term on the right side of (3.17), a lateral boundary integral, we multiply (2.3) by $(x_k - z\delta_{3k}) u_{i,k}$ and integrate it over R_z:

$$\int_{R_z} (x_k - z\delta_{3k}) u_{i,k} [(u_{i,j} - u_{j,i})_{,j} - u_{i,t} - J_1 u_i - J_2 |\mathbf{u}| u_i - p_{,i}] dx = 0,$$

where δ_{ij} is the Kronecker delta symbol. Integration by parts results in

$$\begin{aligned}
& \int_{\partial R_z} (x_k - z\delta_{3k}) u_{i,k} (u_{i,j} - u_{j,i}) n_j ds \\
& \quad - \frac{1}{2} \int_{\partial R_z} (x_k - z\delta_{3k}) u_{i,k} (u_{i,j} - u_{j,i}) n_k ds \\
& \quad + \frac{1}{2} \int_{R_z} (u_{i,j} - u_{j,i}) u_{i,j} dx - \int_{R_z} (x_k - z\delta_{3k}) u_{i,k} u_{i,t} dx \\
& \quad + \frac{1}{2} J_1 \int_{R_z} u_i u_i dx + \frac{1}{3} J_2 \int_{R_z} (u_i u_i)^{\frac{3}{2}} dx - \int_{\partial R_z} (x_k - z\delta_{3k}) u_{i,k} p n_i ds = 0.
\end{aligned} \tag{3.18}$$

Since $u_{i,k} n_i = u_{i,i} n_k = 0$ on ∂D_ξ for $\xi \geq 0$, we rewrite the last term on the right side as

$$\int_{\partial R_z} (x_k - z\delta_{3k}) u_{i,k} p n_i ds = \int_{D_z} x_\alpha u_{3,\alpha} p dA.$$

Moreover, the first two integrals on the lateral boundary can be combined so that

$$\begin{aligned}
& \int_z^\infty \int_{\partial D_\xi} (x_k - z\delta_{3k}) u_{i,k} (u_{i,j} - u_{j,i}) n_j ds d\xi \\
& \qquad - \frac{1}{2} \int_z^\infty \int_{\partial D_\xi} (x_k - z\delta_{3k}) u_{i,k} (u_{i,j} - u_{j,i}) n_k ds d\xi \\
& \quad = \frac{1}{2} \int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha u_{i,j} (u_{i,j} - u_{j,i}) ds d\xi.
\end{aligned}$$

Consequently, (3.18) can be rewritten as

$$\begin{aligned}\frac{1}{2}\int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha u_{i,j}(u_{i,j}-u_{j,i})ds d\xi + \frac{1}{2}J_1\int_{R_z} u_i u_i dx \\ + \frac{1}{3}J_2\int_{R_z}(u_i u_i)^{\frac{3}{2}}dx = \int_{D_z} x_\alpha u_{\beta,\alpha}(u_{\beta,3}-u_{3,\beta})dA \\ -\frac{1}{2}\int_{R_z}(u_{i,j}-u_{j,i})u_{i,j}dx + \int_{R_z}(x_3-z)u_{i,3}u_{i,t}dx \\ + \int_{R_z} x_\alpha u_{i,\alpha}u_{i,t}dx + \int_{D_z} x_\alpha u_{3,\alpha} p dA. \qquad (3.19)\end{aligned}$$

We define $\bar{p}$ to be the mean value of p over D_z, i.e.,

$$\bar{p} = \frac{1}{|D_z|}\int_{D_z} p dA,$$

where $|D_z| = |D_0|$ is the measure of D_z. We note that

$$\int_{D_z} x_\alpha u_{3,\alpha} p dA = \int_{D_z} x_\alpha u_{3,\alpha}(p-\bar{p})dA$$

since

$$\int_{D_z} x_\alpha u_{3,\alpha} p dA = -2\int_{D_z} u_3 dA = 0.$$

Using the Schwarz inequality, it follows from (3.19) that

$$\begin{aligned}\frac{h_0}{2}\int_z^\infty \int_{\partial D_\xi} u_{i,j}(u_{i,j}-u_{j,i})ds d\xi + \frac{1}{2}J_1\int_{R_z} u_i u_i dx + \frac{1}{3}J_2\int_{R_z}(u_i u_i)^{\frac{3}{2}}dx \\ \le d\Big(\int_{D_z} u_{\beta,\alpha}u_{\beta,\alpha}\,dA\Big)^{\frac{1}{2}}\Big(\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})\,dA\Big)^{\frac{1}{2}} \\ + d\Big(\int_{R_z} u_{i,\alpha}u_{i,\alpha}\,dx\Big)^{\frac{1}{2}}\Big(\int_{R_z} u_{i,t}u_{i,t}\,dxd\Big)^{\frac{1}{2}} \\ + \Big(\int_{R_z}(\xi-z)u_{i,3}u_{i,3}\,dx\Big)^{\frac{1}{2}}\Big(\int_{R_z}(\xi-z)u_{i,t}u_{i,t}\,dx\Big)^{\frac{1}{2}} \\ + d\Big(\int_{D_z} u_{3,\alpha}u_{3,\alpha}\,dA\Big)^{\frac{1}{2}}\Big(\int_{D_z}(p-\bar{p})^2\,dA\Big)^{\frac{1}{2}}, \qquad (3.20)\end{aligned}$$

where $h_0 = \min\{x_\alpha n_\alpha\}$ on ∂D_ξ and $d = \text{diameter}\, D_0$. To estimate the last integral on the right side of (3.20), we consider the boundary value problem

$$\Delta\Psi = 0 \quad \text{in } R_z, \qquad \frac{\partial\Psi}{\partial\mathbf{n}} = 0 \quad \text{on } \partial D_\xi, \qquad \frac{\partial\Psi}{\partial\mathbf{n}} = p-\bar{p} \quad \text{in } D_z, \qquad (3.21)$$

where $\xi \ge z \ge 0$. Since $\int_{D_z}(p-\bar{p})dA = 0$, the necessary condition for the existence of a solution Ψ is satisfied. We choose $\lim_{z\to\infty}\int_{D_z}\Psi dA = 0$ such that Ψ is uniquely determined. Using the auxiliary function Ψ, we observe that

$$\begin{aligned}\int_{D_z}(p-\bar{p})^2\,dA = \int_{\partial R_z}(p-\bar{p})\frac{\partial\Psi}{\partial\mathbf{n}}\,ds = \int_{R_z}(p-\bar{p})_{,i}\Psi_{,i}\,dx \\ = \int_{R_z}\Psi_{,i}[(u_{i,j}-u_{j,i})_{,j} - u_{i,t} - J_1 u_i - J_2|\mathbf{u}|u_i]dx.\end{aligned}$$

Then in a manner similar to the derivation of (3.16), we can write

$$\begin{aligned}\int_{D_z}(p-\bar p)^2\,dA \le &\Big(\int_{R_z}\Psi_{,i}\Psi_{,i}\,dx\Big)^{\frac12}\Big(\int_{R_z}u_{i,t}u_{i,t}\,dx\Big)^{\frac12}\\ &+\Big(\int_{D_z}\Psi_{,\alpha}\Psi_{,\alpha}\,dA\Big)^{\frac12}\Big(\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})\,dA\Big)^{\frac12}\\ &+\Big(\int_z^\infty\int_{\partial D_\xi}|\mathrm{grad}_s\Psi|^2\,dsd\xi\Big)^{\frac12}\Big(\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}\,dsd\xi\Big)^{\frac12}\\ &+J_1\Big(\int_{R_z}\Psi_{,i}\Psi_{,i}\,dx\Big)^{\frac12}\Big(\int_{R_z}u_iu_i\,dx\Big)^{\frac12}\\ &+J_2\Big(\int_{R_z}\Psi_{,i}\Psi_{,i}\,dx\Big)^{\frac12}\Big(\int_{R_z}(u_iu_i)^2\,dx\Big)^{\frac12}.\end{aligned}\tag{3.22}$$

The use of (A.9)–(A.11) in the appendix with $g = p - \bar p$ to bound the integrals involving the function Ψ results in an estimate for the integral of the function $p - \bar p$.

$$\begin{aligned}\Big(\int_{D_z}(p-\bar p)^2\,dA\Big)^{\frac12} \le &\Big(\frac{1}{\sqrt{\mu_2}}\int_{R_z}u_{i,t}u_{i,\,t}\,dx\Big)^{\frac12}\\ &+\Big(\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})\,dA\Big)^{\frac12}\\ &+\Big(\frac{2}{h_0}\Big[d+\frac{1}{\sqrt{\mu_2}}\Big]\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}\,dsd\xi\Big)^{\frac12}\\ &+J_1\Big(\frac{1}{\sqrt{\mu_2}}\int_{R_z}u_iu_i\,dx\Big)^{\frac12}+J_2\Big(\frac{1}{\sqrt{\mu_2}}\int_{R_z}(u_iu_i)^2\,dx\Big)^{\frac12}.\end{aligned}\tag{3.23}$$

Substituting (3.23) into (3.20) and using the weighted AG inequalities, we have

$$\begin{aligned}&\frac{h_0}{2}\int_z^\infty\int_{\partial D_\xi}u_{i,j}(u_{i,j}-u_{j,i})dsd\xi\\ &\qquad+\frac12 J_1\int_{R_z}u_iu_idx+\frac13 J_2\int_{R_z}(u_iu_i)^{\frac32}dx\\ &\quad\le\frac{d\gamma_2}{2}\int_{R_z}u_{i,\alpha}u_{i,\alpha}\,dx+\Big(\frac{d}{2\gamma_2}+\frac{d}{2\gamma_4}\Big)\int_{R_z}u_{i,t}u_{i,t}\,dx\\ &\qquad+\frac{\gamma_3}{2}\int_{R_z}(\xi-z)u_{i,3}u_{i,3}\,dx+\frac{1}{2\gamma_3}\int_{R_z}(\xi-z)u_{i,t}u_{i,t}\,dx\\ &\qquad+\Big(\frac{d\gamma_5}{2}+\frac{d^2\gamma_6}{h_0}\Big[d+\frac{1}{\sqrt{\mu_2}}\Big]+\frac{d\gamma_4+d\gamma_7J_1+d^2\gamma_8J_2}{2\sqrt{\mu_2}}\Big)\times\int_{D_z}u_{3,\alpha}u_{3,\alpha}\,dA\\ &\qquad+\frac{d\gamma_1}{2}\int_{D_z}u_{\beta,\alpha}u_{\beta,\alpha}\,dA\\ &\qquad+\Big(\frac{d}{2\gamma_1}+\frac{d}{2\gamma_5}\Big)\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})\,dA\\ &\qquad+\frac{1}{2\gamma_6}\int_z^\infty\int_{\partial D_\xi}u_{i,j}(u_{i,j}-u_{j,i})dsd\xi+\frac{J_1d}{2\gamma_7}\int_{R_z}u_iu_idx+\frac{J_2}{2\gamma_8}\int_{R_z}(u_iu_i)^2dx.\end{aligned}\tag{3.24}$$

We recall that $k = 1/\lambda_1$ and choose

$$\gamma_1 = \gamma_5 = \frac{d\sqrt{\lambda_1}}{2}, \qquad \gamma_2 = \gamma_4 = \sqrt{2\lambda_1}, \qquad \gamma_3 = \sqrt{\lambda_1},$$
$$\gamma_6 = \frac{2}{h_0}, \qquad \gamma_7 = \frac{1}{\sqrt{2\lambda_1}}, \qquad \gamma_8 = \frac{1}{\sqrt{\lambda_1}}.$$

Integrating (3.24) with respect to τ, we thus have

$$\begin{aligned}
&\frac{h_0}{4}\int_0^t \int_z^\infty \int_{\partial R_z} u_{i,j}(u_{i,j} - u_{j,i})\,ds\,d\xi\,d\tau \\
&\qquad \le \frac{\sqrt{\lambda_1}}{2}E + \sqrt{\lambda_1}\left[\frac{d\sqrt{2}}{2} + \frac{J_2\Omega^{\frac{1}{3}}}{2}\max_{0\le\tau\le t}\left(\int_{R_z}(u_iu_i)^{\frac{3}{2}}\,dx\right)^{\frac{2}{3}}\right]\left(-\frac{\partial E}{\partial z}\right) \\
&\qquad\qquad + \frac{2}{\sqrt{\lambda_1}}\int_0^t\int_{D_z}(u_{\alpha,3} - u_{3,\alpha})(u_{\alpha,3} - u_{3,\alpha})\,dA\,d\tau \\
&\qquad\qquad + \frac{d^2\sqrt{\lambda_1}}{4}\int_0^t\int_{D_z} u_{\beta,\alpha}u_{\beta,\alpha}\,dA\,d\tau + \frac{r_1}{\sqrt{\lambda_1}}\int_0^t\int_{D_z} u_{3,\alpha}u_{3,\alpha}\,dA\,d\tau,
\end{aligned} \tag{3.25}$$

where

$$r_1 = \frac{d\sqrt{\lambda_1}}{2}\left(\frac{d\sqrt{\lambda_1}}{2} + \frac{4d}{h_0^2}\left[d + \frac{1}{\sqrt{\mu_2}}\right] + \frac{\sqrt{2\lambda_1}}{\sqrt{\mu_2}} + \frac{J_1}{\sqrt{2\lambda_1\mu_2}} + \frac{dJ_2}{\sqrt{\lambda_1\mu_2}}\right). \tag{3.26}$$

Now we return to (3.17) and use (3.25) to obtain

$$\begin{aligned}
I_4 &\le \frac{1}{2}E + r_2\left(-\frac{\partial E}{\partial z}\right) + \frac{d^2}{4}\int_0^t\int_{D_z} u_{\beta,\alpha}u_{\beta,\alpha}\,dA\,d\tau \\
&\qquad + \frac{3}{\lambda_1}(1+\eta)\int_0^t\int_{D_z} u_{\alpha,3}u_{\alpha,3}\,dA\,d\tau \\
&\qquad + \frac{3}{\lambda_1}\left(1 + \frac{1}{\eta} + \frac{r_1}{3}\right)\int_0^t\int_{D_z} u_{3,\alpha}u_{3,\alpha}\,dA\,d\tau,
\end{aligned} \tag{3.27}$$

where

$$\begin{aligned}
r_2 &= \frac{3\sqrt{\lambda_1} + J_1^{\frac{1}{2}} + J_2}{2\sqrt{\lambda_1\mu_2}} + \frac{2}{h_0^2\sqrt{\lambda_1}}\left[\frac{1}{\mu_2} + \frac{d^2}{4}\right] + \frac{d\sqrt{2}}{2} \\
&\qquad + \left(\frac{J_2\Omega^{\frac{1}{3}}}{2} + \frac{J_2\Omega^{\frac{1}{3}}}{2\sqrt{\lambda_1\mu_2}}\right)\max_{0\le\tau\le t}\left(\int_{R_z}(u_iu_i)^{\frac{3}{2}}(\tau)\,dx\right)^{\frac{2}{3}}.
\end{aligned} \tag{3.28}$$

We choose η to be the positive root of the quadratic equation $\eta^2 - (r_1/3)\eta - 1 = 0$ so that the coefficients of the last two integrals of the right side of (3.27) are equal. Consequently, we can rewrite (3.27) as

$$I_4 \le \frac{1}{2}E + r_2\left(-\frac{\partial E}{\partial z}\right) + r_3\frac{\partial^2 E}{\partial z^2}, \tag{3.29}$$

where

$$r_3 = \max\left\{\frac{d^2}{4}, \frac{3}{\lambda_1}(1+\eta)\right\}.$$

Combining all of the results in (3.4), (3.14), and (3.29), we arrive at the second-order differential inequality

$$E \leq K_1 \frac{\partial^2 E}{\partial z^2} - K_2 \frac{\partial E}{\partial z}, \tag{3.30}$$

where

$$K_1 = \frac{\sqrt{C}}{\lambda_1} + 2r_3,$$

$$\begin{aligned} K_2 = {} & 2\left(\frac{1+\sqrt{C}}{\sqrt{\lambda_1}}\right) + \frac{3\sqrt{\lambda_1} + J_1^{\frac{1}{2}} + J_2}{\sqrt{\lambda_1 \mu_2}} + \frac{4}{h_0^2 \sqrt{\lambda_1}}\left[\frac{1}{\mu_2} + \frac{d^2}{4}\right] + d\sqrt{2} \\ & + \frac{2J_2 \Omega^{\frac{1}{6}} C^{\frac{1}{2}}}{\lambda_1} \max_{0 \leq \tau \leq t} \left(\int_{R_z} (u_i u_i)^{\frac{3}{2}}(\tau)\, dx \right)^{\frac{1}{3}} \\ & + \left(J_2 \Omega^{\frac{1}{3}} + \frac{J_2 \Omega^{\frac{1}{3}}}{\sqrt{\lambda_1 \mu_2}} \right) \max_{0 \leq \tau \leq t} \left(\int_{R_z} (u_i u_i)^{\frac{3}{2}}(\tau)\, dx \right)^{\frac{2}{3}}. \end{aligned} \tag{3.31}$$

We may rewrite (3.30) as

$$\frac{\partial}{\partial z}\left\{ e^{-k_1 z}\left(\frac{\partial E}{\partial z} + k_2 E \right) \right\} \geq 0, \tag{3.32}$$

where

$$k_1 = \frac{K_2}{2K_1} + \frac{1}{2}\sqrt{\frac{K_2^2}{K_1^2} + \frac{4}{K_1}}, \qquad k_2 = -\frac{K_2}{2K_1} + \frac{1}{2}\sqrt{\frac{K_2^2}{K_1^2} + \frac{4}{K_1}}.$$

Integration of (3.32) from z to ∞ leads to

$$\frac{\partial E}{\partial z} + k_2 E \leq 0,$$

and hence

$$E(z,t) \leq E(0,t) e^{-k_2 z}. \tag{3.33}$$

Direct calculation shows that

$$\frac{\partial k_2}{\partial K_1} = \frac{1}{2}\frac{1}{K_2}\left(\frac{K_1^2}{K_2^2} + \frac{4}{K_2} \right)^{-\frac{1}{2}} \left(-\left(\frac{K_1^2}{K_2^2} + \frac{4}{K_2} \right)^{\frac{1}{2}} + \frac{K_1}{K_2} \right) < 0 \tag{3.34}$$

and

$$\frac{\partial k_2}{\partial K_2} = \frac{1}{2}\frac{K_1}{K_2^2}\left(\frac{K_1^2}{K_2^2} + \frac{4}{K_2} \right)^{-\frac{1}{2}} \left(\left(\frac{K_1^2}{K_2^2} + \frac{4}{K_2} \right)^{\frac{1}{2}} - \left(\frac{K_1}{K_2} + \frac{2}{K_1} \right) \right) < 0. \tag{3.35}$$

Thus we find that

$$\frac{\partial k_2}{\partial J_1} = \frac{\partial k_2}{\partial K_1}\frac{\partial K_1}{\partial J_1} + \frac{\partial k_2}{\partial K_2}\frac{\partial K_2}{\partial J_1} < 0, \qquad \frac{\partial k_2}{\partial J_2} = \frac{\partial k_2}{\partial K_1}\frac{\partial K_1}{\partial J_2} + \frac{\partial k_2}{\partial K_2}\frac{\partial K_2}{\partial J_2} < 0. \tag{3.36}$$

Therefore, the decay rate k_2 is a decreasing function of J_1 and J_2. In the physical sense, this suggests that when Stokes flow is compared to flow in a porous medium, the presence of the porous material as well as the inclusion of the inertial drag force each reduces the spatial decay rate of the end effects.

4. Determination of the bounds. As indicated in the previous section, the definition of the constant K_2 involves the quantity $\max_{0\le\tau\le t}\int_{R_z}(u_iu_i)^{\frac{3}{2}}\,dx$. One of the important aspects in our analysis is that we would like our estimate to depend only on the known data, which for the present problem are the physical parameters J_1 and J_2, the boundary data, and the geometry of the domain. We therefore need to find a bound for the expression $\max_{0\le\tau\le t}\int_R(u_iu_i)^{\frac{3}{2}}\,dx$ as well as for the total weighted energy $E\ (0,\,t)$. Modeled on the analysis used by Ames and Payne [2] in their investigation of decay estimates for the solution of the steady Navier–Stokes equation, we shall consider the solution of the linearized problem, namely the Brinkman flow

$$w_{i,t}=\Delta w_i-J_1w_i-q_{,i}\quad\text{in } R\times(0,\infty),\tag{4.1}$$

$$w_{i,i}=0\quad\text{in } \bar{R}\times(0,\infty),\tag{4.2}$$

$$w_i=u_i\quad\text{on } \partial R\times(0,\infty),\tag{4.3}$$

$$w_i=0\quad\text{in } R\times\{0\},\tag{4.4}$$

$$w_{i},\,w_{i,j},\,w_{i,t},\,q=o(x_3^{-1})\quad\text{uniformly in } x_{1,}\,x_2,\,t\qquad\text{as } x_3\to\infty.\tag{4.5}$$

In what follows, we first compare the solution u_i of problem (2.3)–(2.8), the Brinkman–Forchheimer flow, with the solution w_i of problem (4.1)–(4.5), the Brinkman flow. Next, we explain how one can bound the integrals containing the solution w_i and its derivatives in section 4.2.

4.1. Bound for $\max_{0\le\tau\le t}\int_R(u_iu_i)^{\frac{3}{2}}\,dx$.

To relate the solution u_i of system (2.3)–(2.8) to the solution w_i of system (4.1)–(4.5), we define $v_i=u_i-w_i$ and $s=p-q$. Then the initial-boundary problem governing the differential field is

$$v_{i,t}=\Delta v_i-J_1v_i-J_2|\mathbf{u}|u_i-s_{,i}\quad\text{in } R\times(0,\infty),\tag{4.6}$$

$$v_{i,i}=0\quad\text{in } \bar{R}\times(0,\infty),\tag{4.7}$$

$$v_i=0\quad\text{on } \partial R\times(0,\infty),\tag{4.8}$$

$$v_i=0\quad\text{in } R\times\{0\},\tag{4.9}$$

$$v_{i},\,v_{i,j},\,v_{i,t},\,s=o(x_3^{-1})\quad\text{uniformly in } x_{1,}\,x_2,\,t\qquad\text{as } x_3\to\infty.\tag{4.10}$$

On integration by parts and dropping one term which is negative, we first note for any $0\le t_1\le t$ that

$$\begin{aligned}
0&\le\int_0^{t_1}\int_R v_{i,\tau}v_{i,\tau}\,dxd\tau\\
&\le\int_0^{t_1}\int_R v_{i,\tau}[\Delta v_i-J_1v_i-J_2|\mathbf{u}|u_i-s_{,i}]\,dxd\tau\\
&\le-\frac{1}{2}\int_0^{t_1}\int_R(v_{i,j}v_{i,j})_{,\tau}\,dxd\tau-J_2\int_0^{t_1}\int_R v_{i,\tau}|\mathbf{u}|u_i\,dxd\tau\\
&\le-\frac{1}{2}\int_0^{t_1}\int_R(v_{i,j}v_{i,j})_{,\tau}\,dxd\tau-J_2\int_0^{t_1}\int_R|\mathbf{u}|u_i(u_{i,\tau}-w_{i,\tau})\,dxd\tau.
\end{aligned}\tag{4.11}$$

This yields that

$$\begin{aligned}
&\frac{1}{2}\int_0^{t_1}\int_R(v_{i,j}v_{i,j})_{,\tau}\,dxd\tau+J_2\int_0^{t_1}\int_R|\mathbf{u}|u_iu_{i,\tau}\,dxd\tau+\int_0^{t_1}\int_R v_{i,\tau}v_{i,\tau}\,dxd\tau\\
&\le J_2\int_0^{t_1}\int_R|\mathbf{u}|u_iw_{i,\tau}\,dxd\tau.
\end{aligned}\tag{4.12}$$

For the above inequality, we integrate two terms on the left side with respect to τ and use the initial condition (4.9) and then apply the Schwarz inequality to the integral on the right side. It is then true for any $0 \leq t_1 \leq t$ that

$$\begin{aligned}&\frac{1}{2}\int_R v_{i,j}v_{i,j}(t_1)\,dx + \frac{1}{3}J_2\int_R (u_iu_i)^{\frac{3}{2}}\,(t_1)dx + \int_0^{t_1}\int_R v_{i,\tau}v_{i,\tau}\,dxd\tau \\ &\leq J_2\left(\int_0^t\int_R w_{i,\tau}w_{i,\tau}\,dxd\tau\right)^{\frac{1}{2}}\left(\int_0^t\int_R (u_iu_i)^2\,dxd\tau\right)^{\frac{1}{2}}. \end{aligned}\tag{4.13}$$

We recall the result in (3.12) which provides that

$$\begin{aligned}&\int_0^t\int_R (u_iu_i)^2\,dxd\tau \\ &\leq \Omega^{\frac{1}{3}}\left(\max_{0\leq\tau\leq t}\int_R (u_iu_i)^{\frac{3}{2}}(\tau)\,dx\right)^{\frac{2}{3}}\int_0^t\int_R u_{i,j}u_{i,j}\,dxd\tau. \end{aligned}\tag{4.14}$$

We conclude immediately from (4.13) and (4.14) that

$$\begin{aligned}&\left(\max_{0\leq\tau\leq t}\int_R (u_iu_i)^{\frac{3}{2}}(\tau)\,dx\right)^{\frac{2}{3}} \\ &\leq 3\Omega^{\frac{1}{6}}\left(\int_0^t\int_R w_{i,\tau}w_{i,\tau}\,dxd\tau\right)^{\frac{1}{2}}\left(\int_0^t\int_R u_{i,j}u_{i,j}\,dxd\tau\right)^{\frac{1}{2}}. \end{aligned}\tag{4.15}$$

Using (4.15), we have that the term $\max_{0\leq\tau\leq t}\int_R (u_iu_i)^{\frac{3}{2}}(\tau)\,dx$ can be bounded if the two integrals on the right side of (4.15) are bounded. We shall discuss the bounds of the integral $\int_0^t\int_R w_{i,\tau}w_{i,\tau}\,dxd\tau$ in section 4.2. For the integral $\int_0^t\int_R u_{i,j}u_{i,j}\,dxd\tau$, we want to show that it can be bounded by two integrals of the solution w_i.

On integration by parts and dropping three negative terms, we observe that

$$\begin{aligned}\int_0^t\int_R v_{i,j}v_{i,j}\,dxd\tau &= -\int_0^t\int_R v_iv_{i,jj}\,dxd\tau \\ &= -\int_0^t\int_R v_i[v_{,\tau} + J_1v_i + J_2|\mathbf{u}|u_i + s_{,i}]\,dxd\tau \\ &\leq -J_2\int_0^t\int_R |\mathbf{u}|v_i(v_i+w_i)\,dxd\tau \\ &\leq \frac{J_2}{4}\int_0^t\int_R |\mathbf{u}|w_iw_i\,dxd\tau \\ &\leq \frac{J_2}{4}\int_0^t\left(\int_R (u_iu_i)^{\frac{3}{2}}\,dx\right)^{\frac{1}{3}}\left(\int_R (w_iw_i)^{\frac{3}{2}}\,dx\right)^{\frac{2}{3}}d\tau \\ &\leq \frac{J_2}{4}\max_{0\leq\tau\leq t}\left(\int_R (u_iu_i)^{\frac{3}{2}}\,dx\right)^{\frac{1}{3}}\left\{\int_0^t\left(\int_R (w_iw_i)^{\frac{3}{2}}\,dx\right)^{\frac{2}{3}}d\tau\right\}. \end{aligned}\tag{4.16}$$

Setting $w = (w_iw_i)^{\frac{1}{2}}$ in (A.4), we apply the Sobolev inequality to obtain that

$$\begin{aligned}\int_R (w_iw_i)^2\,dx &= \int_0^\infty\int_{D_\xi}(w_iw_i)^2\,dAd\xi \\ &\leq \frac{1}{2}\int_0^\infty\left[\int_{D_\xi} w_iw_i\,dA\right]\left[\int_{D_\xi} w_{i,\beta}w_{i,\beta}\,dA\right]d\xi. \end{aligned}\tag{4.17}$$

The divergence theorem and boundary conditions for w_i on ∂D_z yield, for each ξ in $z \le \xi < \infty$,

$$\begin{aligned}\int_{D_\xi} w_i w_i \, dA &= -2\int_z^\infty \int_{D_\xi} w_i w_{i,\xi} \, dA d\xi \\ &\le 2\Big(\int_{R_z} w_i w_i \, dx\Big)^{\frac{1}{2}} \Big(\int_{R_z} w_{i,3} w_{i,3} \, dx\Big)^{\frac{1}{2}} \\ &\le 2\Big(\int_{R} w_i w_i \, dx\Big)^{\frac{1}{2}} \Big(\int_{R} w_{i,3} w_{i,3} \, dx\Big)^{\frac{1}{2}}. \end{aligned} \tag{4.18}$$

Using (4.17), (4.18), and (A.1), we obtain

$$\begin{aligned}\int_R (w_i w_i)^2 \, dx &\le \Big(\int_R w_i w_i \, dx\Big)^{\frac{1}{2}} \Big(\int_R w_{i,j} w_{i,j} \, dx\Big)^{\frac{3}{2}} \\ &\le \frac{1}{\sqrt{\lambda_1}} \Big(\int_R w_{i,j} w_{i,j} \, dx\Big)^2. \end{aligned} \tag{4.19}$$

Furthermore, we have

$$\begin{aligned}\Big[\int_R (w_i w_i)^{\frac{3}{2}} \, dx\Big]^{\frac{2}{3}} &\le \Big[\Big(\int_R (w_i w_i) \, dx\Big)^{\frac{1}{2}} \Big(\int_R (w_i w_i)^2 \, dx\Big)^{\frac{1}{2}}\Big]^{\frac{2}{3}} \\ &\le \frac{1}{\sqrt{\lambda_1}} \int_R w_{i,j} w_{i,j} \, dx. \end{aligned} \tag{4.20}$$

Substituting (4.15) and (4.20) in (4.16), we thus obtain

$$\begin{aligned}&\int_0^t \int_R v_{i,j} v_{i,j} \, dx d\tau \\ &\quad \le \frac{J_2}{4\lambda_1{}^{\frac{1}{2}}} \big(\max_{0\le\tau\le t} \int_R (u_i u_i)^{\frac{3}{2}} \, dx\big)^{\frac{1}{3}} \int_0^t \int_R w_{i,j} w_{i,j} \, dx d\tau \\ &\quad \le \frac{3^{\frac{1}{2}} \Omega^{\frac{1}{12}} J_2}{4\lambda_1{}^{\frac{1}{2}}} \Big(\int_0^t \int_R w_{i,j} w_{i,j} \, dx d\tau\Big) \Big(\int_0^t \int_R w_{i,\tau} w_{i,\tau} \, dx d\tau\Big)^{\frac{1}{4}} \\ &\qquad \times \Big(\int_0^t \int_R u_{i,j} u_{i,j} \, dx d\tau\Big)^{\frac{1}{4}}. \end{aligned} \tag{4.21}$$

Note that

$$\int_0^t \int_R u_{i,j} u_{i,j} \, dx d\tau \le 2\int_0^t \int_R v_{i,j} v_{i,j} \, dx d\tau + 2\int_0^t \int_R w_{i,j} w_{i,j} \, dx d\tau, \tag{4.22}$$

and then one has

$$\int_0^t \int_R u_{i,j} u_{i,j} \, dx d\tau \le F\Big(\Big[\int_0^t \int_R w_{i,j} w_{i,j} \, dx d\tau\Big], \Big[\int_0^t \int_R w_{i,\tau} w_{i,\tau} \, dx d\tau\Big]\Big), \tag{4.23}$$

where F is a computable function. In fact, this can be seen from the following arguments. Let $A = \int_0^t \int_R w_{i,j} w_{i,j} \, dx d\tau$, $B = \int_0^t \int_R w_{i,\tau} w_{i,\tau} \, dx d\tau$, and $Y = \int_0^t \int_R u_{i,j} u_{i,j} \, dx d\tau$. From (4.21) and (4.22), it follows that

$$Y \le \Gamma A B^{\frac{1}{4}} Y^{\frac{1}{4}} + 2A, \tag{4.24}$$

where $\Gamma = 3^{1/2}\Omega^{1/12}J_2/2\lambda_1^{1/2}$. Without loss of generality, we can assume that $Y > 1$; otherwise $F(a,b) = 1$. With this assumption, it is easy to see that

$$Y \leq \Gamma A B^{\frac{1}{4}} Y^{\frac{1}{2}} + 2A. \tag{4.25}$$

Solving inequality (4.25), we have

$$Y \leq \frac{1}{4}\left(\Gamma A B^{\frac{1}{4}} + \sqrt{\Gamma^2 A^2 B^{\frac{1}{2}} + 8A}\right)^2,$$

which proves that (4.23) holds provided that $F(A,B) = \max\{1, (1/4)(\Gamma A B^{\frac{1}{4}} + \sqrt{\Gamma^2 A^2 B^{\frac{1}{2}} + 8A})^2\}$.

From (4.13)–(4.15), it also follows that

$$\int_0^t \int_R v_{i,\tau}v_{i,\tau}\,dxd\tau \leq J_2\sqrt{3}\Omega^{\frac{1}{4}}\left(\int_0^t \int_R w_{i,\tau}w_{i,\tau}\,dxd\tau\right)^{\frac{3}{4}}\left(\int_0^t \int_R u_{i,j}u_{i,j}\,dxd\tau\right)^{\frac{3}{4}},$$

and then by using the elementary inequality $\frac{1}{2}a^2 - b^2 \leq (a-b)^2$, it follows that

$$\begin{aligned}\int_0^t \int_R u_{i,\tau}u_{i,\tau}\,dxd\tau \leq{}& 2\sqrt{3}J_2\Omega^{\frac{1}{4}}\left(\int_0^t \int_R w_{i,\tau}w_{i,\tau}\,dxd\tau\right)^{\frac{3}{4}}\left(\int_0^t \int_R u_{i,j}u_{i,j}\,dxd\tau\right)^{\frac{3}{4}}\\ &+ 2\int_0^t \int_R w_{i,\tau}w_{i,\tau}\,dxd\tau.\end{aligned} \tag{4.26}$$

Inequality (4.26) together with (4.23) indicates that the integral $\int_0^t \int_R u_{i,\tau}u_{i,\tau}\,dxd\tau$ is also bounded by the integrals $\int_0^t \int_R w_{i,j}w_{i,j}\,dxd\tau$ and $\int_0^t \int_R w_{i,\tau}w_{i,\tau}\,dxd\tau$.

Although the above arguments are similar to those used in [2] to compare the solution of the nonlinear equation with the solution of the linearized equation, the conclusions reached here do not require restrictions because of the above argument. In studying the Navier–Stokes pipe flow, conditions were imposed by Ames and Payne [2] in order to bound the energy expression, which effectively said that the decay is ensured only for flows with sufficiently large viscosity coefficients (or small Reynolds' number) and for flows whose data are suitably restricted.

4.2. Bounds for the Brinkman flow. In this section, we focus on finding the bounds for the two integrals $\int_0^t \int_R w_{i,j}w_{i,j}v\,dxd\tau$ and $\int_0^t \int_R w_{i,\tau}w_{i,\tau}\,dxd\tau$ in terms of the physical parameters, boundary data, and geometry of the domain. We use E_1, $\partial E_1/\partial z$, and $\partial^2 E_1/\partial z^2$ to denote the weighted energy integral and its first and second derivatives for the solution w_i of system (4.1)–(4.5). They are defined in the same manner as in (2.9)–(2.11), replacing u_i by w_i. For simplicity, we shall not determine the desired bound explicitly but indicate only how the bounds can be found from the known quantities. Here we let ϵ_i denote positive constants which may be chosen arbitrarily small and c_i denote computable constants that may depend on ϵ_i.

Since the solution w_i and the solution u_i have the same initial and boundary data, on integration by parts, we observe that

$$\begin{aligned}\int_0^t\int_R w_{i,j}w_{i,j}\,dxd\tau &= -\int_0^t\int_{D_0} w_iw_{i,3}\,dAd\tau - \int_0^t\int_R w_iw_{i,jj}\,dxd\tau\\ &= -\int_0^t\int_{D_0} w_iw_{i,3}\,dAd\tau - \int_0^t\int_R w_i\big(w_{i,\tau}+J_1w_i+q_{,i}\big)\,dxd\tau\\ &\le -J_1\int_0^t\int_R w_iw_i\,dxd\tau - \int_0^t\int_{D_0} f_3f_{\alpha,\alpha}\,dAd\tau\\ &\quad -\int_0^t\int_{D_0} f_3q\,dAd\tau - \int_0^t\int_{D_0} f_\alpha w_{\alpha,3}\,dAd\tau \end{aligned}\tag{4.27}$$

and

$$\begin{aligned}\int_0^t\int_R w_{i,\tau}w_{i,\tau}\,dxd\tau &= \int_0^t\int_R w_{i,\tau}\big(w_{i,jj}-J_1w_i-q_{,i}\big)\,dxd\tau\\ &\le -\int_0^t\int_{D_0} f_{i,\tau}w_{i,3}\,dAd\tau + \int_0^t\int_{D_0} f_{3,\tau}q\,dAd\tau\\ &= \int_0^t\int_{D_0} f_{3,\tau}f_{\alpha,\alpha}\,dAd\tau - \int_0^t\int_{D_0} f_{\alpha,\tau}w_{\alpha,3}\,dAd\tau\\ &\quad +\int_0^t\int_{D_0} f_{3,\tau}q\,dAd\tau. \end{aligned}\tag{4.28}$$

Thus (4.27) and (4.28) imply that

$$\begin{aligned}-\frac{\partial E_1}{\partial z}(0,\,t) &= \int_0^t\int_R \big(J_1w_iw_i+w_{i,j}w_{i,j}+kw_{i,\tau}w_{i,\tau}\big)\,dxd\tau\\ &\le \text{data} + \epsilon_1\int_0^t\int_{D_0} w_{\alpha,3}w_{\alpha,3}\,dAd\tau + \epsilon_2\int_0^t\int_{D_0}(q-\bar q)^2\,dAd\tau, \end{aligned}\tag{4.29}$$

where $\bar q = (1/|D_0|)\int_{D_0} q\,dA$ is the mean value of q over D_0 and "data" refers to constants involving only the given data and parameters. Since equation (4.1) is the linearized equation of (2.3), all of the derivations in the previous sections can be obtained simply by letting $J_2 = 0$. This means that we could write

$$E_1(0,t) \le c_1\frac{\partial^2 E_1}{\partial z^2}(0,t) - c_2\frac{\partial E_1}{\partial z}(0,t).\tag{4.30}$$

Moreover, direct calculation shows that

$$\frac{\partial^2 E_1}{\partial z^2}(0,t) \le \text{data} + \int_0^t\int_{D_0} w_{\alpha,3}w_{\alpha,3}\,dAd\tau.\tag{4.31}$$

Using (3.22) and (3.23), we can derive in the same manner that

$$\begin{aligned}\int_0^t\int_{D_0}(q-\bar q)^2\,dAd\tau &\le \text{data} + c_3\int_0^t\int_{D_0} w_{\alpha,3}w_{\alpha,3}\,dAd\tau\\ &\quad - c_4\frac{\partial E_1}{\partial z}(0,t) + c_5E_1(0,t). \end{aligned}\tag{4.32}$$

To establish a bound for the term $\int_0^t \int_{D_0} w_{\alpha,3} w_{\alpha,3}\, dA d\tau$, we consider the identity

$$\int_0^t \int_R w_{i,3} \left[(w_{i,j} - w_{j,i})_{,j} - J_1 w_i - q_{,i} - w_{i,\tau} \right] dx d\tau = 0.$$

On integration by parts, we are led to

$$\begin{aligned} &- \int_0^t \int_{D_0} w_{i,3}(w_{i,3} - w_{3,i})\ dA d\tau + \frac{1}{2} \int_0^t \int_{D_0} w_{i,j}(w_{i,j} - w_{j,i})\, dA d\tau \\ &- J_1 \int_0^t \int_R w_i w_{i,3}\ dx d\tau + \int_0^t \int_{D_0} w_{3,3} q\ dA d\tau - \int_0^t \int_R w_{i,3} w_{i,\tau}\ dx d\tau = 0. \end{aligned}$$

Since $w_{3,3} = -f_{\alpha,\alpha}$ on D_0, we can deduce from this identity that

$$\int_0^t \int_{D_0} w_{\alpha,3} w_{\alpha,3}\, dA d\tau \le \text{data} + \epsilon_3 \int_0^t \int_{D_0} (q - \bar{q})^2\, dA d\tau - c_6 \frac{\partial E_1}{\partial z}(0,t). \tag{4.33}$$

From (4.32) and (4.33), it follows that

$$(1 - c_3\epsilon_3) \int_0^t \int_{D_0} w_{\alpha,3} w_{\alpha,3}\, dA d\tau \le \text{data} - c_4 \epsilon_3 \frac{\partial E_1}{\partial z}(0,t) + c_5 \epsilon_3 E_1(0,t) \tag{4.34}$$

and

$$(1 - c_3\epsilon_3) \int_0^t \int_{D_0} (q - \bar{q})^2\, dA d\tau \le \text{data} - (c_4 + c_3 c_6) \frac{\partial E_1}{\partial z}(0,t) + c_5 E_1(0,t). \tag{4.35}$$

Using (4.31) and (4.34), we can rewrite (4.30) as

$$\frac{1 - c_3\epsilon_3 - c_1 c_5 \epsilon_3}{1 - c_3 \epsilon_3} E_1(0,t) \le \text{data} - \left(c_2 + \frac{c_1 c_4 \epsilon_3}{1 - c_1 \epsilon_3} \right) \frac{\partial E_1}{\partial z}(0,t). \tag{4.36}$$

With an appropriate choice for ϵ_3, we have

$$E_1(0,t) \le \text{data} - c_7 \frac{\partial E_1}{\partial z}(0,t),$$

and we then derive from (4.34) and (4.35) that

$$\int_0^t \int_{D_0} w_{\alpha,3} w_{\alpha,3}\, dA d\tau \le \text{data} - c_8 \frac{\partial E_1}{\partial z}(0,t), \tag{4.37}$$

$$\int_0^t \int_{D_0} (q - \bar{q})^2\, dA d\tau \le \text{data} - c_9 \frac{\partial E_1}{\partial z}(0,t). \tag{4.38}$$

Substituting (4.37) and (4.38) into (4.29) and choosing ϵ_1 and ϵ_2 appropriately, we conclude that

$$-\frac{\partial E_1}{\partial z}(0,t) \le \text{data}, \tag{4.39}$$

which concludes the derivation.

4.3. Bounds for $\boldsymbol{E(0,t)}$. In light of the results above, the quantity $\max_{0\leq\tau\leq t}\int_R (u_iu_i)^{\frac{3}{2}}$ can be bounded by the known data, so the constants in (3.31) depend only on these data. Following the same arguments used in [1, section 6] or the discussions given in section 4.2, we are able to conclude that the total weighted energy $E(0,t)$ has an upper bound in terms of the physical parameters, the boundary data, and the geometry of domain.

5. The estimates for Darcy flow.

For comparison, we discuss the limiting case when J_1 is very large. It is believed that the flow is governed by Darcy's law if the porous medium is dense. In other words, in the limiting case when J_1 is very large, the term involving the time derivative, the Brinkman term, and the Fochheimer term can be ignored so that the model equation has a simple form.

We assume u_i and p are solutions of the following boundary value problem:

$$J_1u_i = -p_i \quad \text{in } R, \tag{5.1}$$

$$u_{i,i} = 0 \quad \text{in } \bar{R}, \tag{5.2}$$

$$u_in_i = 0 \quad \text{on } \partial R\backslash D_0, \tag{5.3}$$

$$u_3 = f(x_1,x_2) \quad \text{in } \bar{D_0}, \tag{5.4}$$

$$u_i,\, p = o(x_3^{-1}) \quad \text{uniformly in } x_{1,}\, x_2 \quad \text{as } x_3\to\infty. \tag{5.5}$$

Define a weighted energy integral for the solution u_i of (5.1)–(5.5) by

$$E(z) = \int_{R_z} J_1(\xi - z)u_iu_i\,dx, \tag{5.6}$$

and so

$$\frac{\partial E}{\partial z} = -J_1\int_{R_z} u_iu_idx, \tag{5.7}$$

Using (5.1), we rewrite E as

$$E(z) = -\int_{R_z}(\xi - z)u_ip_{,i}\,dx = \int_{R_z} u_3p\,dx. \tag{5.8}$$

Following the same techniques used in dealing with I_3 in section 3 (see (3.6)), we can estimate the integral on the right side of (5.8) in terms of the auxiliary vector function ω_α (see (A.3)):

$$\begin{aligned}\int_{R_z} u_3p\,dx &= \int_{R_z}\omega_{\alpha,\alpha}p\,dx = -\int_{R_z}\omega_\alpha p_{,\alpha}\,dx\\ &= \int_{R_z} J_1\omega_\alpha u_\alpha\,dx\\ &\leq J_1\left(\int_{R_z}\omega_\alpha\omega_\alpha\,dx\right)^{\frac{1}{2}}\left(\int_{R_z}u_\alpha u_\alpha\,dxd\tau\right)^{\frac{1}{2}}\\ &\leq J_1\sqrt{\frac{C}{\lambda_1}}\left(\int_{R_z}(u_3)^2\,dx\right)^{\frac{1}{2}}\left(\int_{R_z}u_\alpha u_\alpha\,dx\right)^{\frac{1}{2}}\\ &\leq \frac{J_1\sqrt{C}}{2\sqrt{\lambda_1}}\int_{R_z}u_iu_i\,dx.\end{aligned} \tag{5.9}$$

From (5.8) and (5.9), it follows that

$$E(z) \leq \frac{\sqrt{C}}{2\sqrt{\lambda_1}}\Big(-\frac{\partial E}{\partial z}\Big) \quad \text{for } z \geq 0 \tag{5.10}$$

and hence

$$E(z) \leq E(0,t)e^{-\frac{\sqrt{C}}{2\sqrt{\lambda_1}}z}. \tag{5.11}$$

Using the same techniques as in section 4, we can find the total energy bound explicitly. In fact, we observe that

$$\begin{aligned}\int_R J_1 u_i u_i \, dx &= -\int_R u_i p_{,i} \, dx = \int_{D_0} u_3 p \, dA \\ &\leq \Big(\int_{D_0} f^2 \, dA\Big)^{\frac{1}{2}} \Big(\int_{D_0} (p-\bar{p})^2 \, dA\Big)^{\frac{1}{2}},\end{aligned} \tag{5.12}$$

where $\bar{p} = \int_{D_0} p \, dA$. In a manner similar to the derivation of (3.23), we have

$$\Big(\int_{D_0} (p-\bar{p})^2 \, dA\Big)^{\frac{1}{2}} \leq J_1 \Big(\int_R u_i u_i \, dx\Big)^{\frac{1}{2}}. \tag{5.13}$$

Combining (5.12) and (5.13), we obtain

$$\int_R J_1 u_i u_i \, dx \leq \sqrt{J_1} \int_{D_0} f^2 \, dA. \tag{5.14}$$

Using (5.10) and (5.14), a bound for the total weighted energy in terms of the physical parameter J_1, the boundary data f and the first eigenvalue λ_1 can be given by

$$E(0) \leq \frac{\sqrt{J_1 C}}{2\sqrt{\lambda_1}} \int_{D_0} f^2 \, dA. \tag{5.15}$$

Appendix. Auxiliary inequalities and functions. Here we record some standard inequalities and the properties of two auxiliary functions that we need in order to establish our estimates in the previous sections.

Let S be a plane domain with boundary ∂S.

1. If $w = 0$ on ∂S, then we have the Poincaré inequality

$$\lambda_1 \int_S w^2 \, dA \leq \int_S w_{,\alpha} w_{,\alpha} \, dA, \tag{A.1}$$

where λ_1 is the smallest positive eigenvalue of

$$\Delta\phi + \lambda\phi = 0 \quad \text{in } S, \qquad \phi = 0 \quad \text{on } \partial S.$$

2. If $\partial w / \partial \mathbf{n} = 0$ on ∂S and $\int_S w \, dA = 0$, then the Wirtinger inequality has the form

$$\mu_2 \int_S w^2 \, dA \leq \int_S w_{,\alpha} w_{,\alpha} \, dA, \tag{A.2}$$

where μ_2 is the smallest positive eigenvalue of

$$\Delta\phi + \mu\phi = 0 \quad \text{in } S, \qquad \frac{\partial\phi}{\partial\mathbf{n}} = 0 \quad \text{on } \partial S, \qquad \int_S \phi \, dA = 0.$$

3. If g is a continuously differentiable function on $\bar{S}$ and $\int_S g\,dA = 0$, then there exists a vector function w_α such that

$$w_{\alpha,\alpha} = g \quad \text{in } S, \qquad w_\alpha = 0 \quad \text{on } \partial S$$

and a positive constant C depending on the geometry of S such that

$$\text{(A.3)} \qquad \int_S w_{\alpha,\beta} w_{\alpha,\beta} \, dA \leq C \int_S (w_{\alpha,\alpha})^2 \, dA.$$

The implication above asserts the existence of a vector function w_α which is, in fact, not unique. We require only the existence of such a vector function in our derivation and not an explicit solution. We refer the reader to [2] for a brief discussion concerning the constant C and to [8] for an explicit upper bound for the optimal C when S is a star-shaped domain.

4. We shall make use of two Sobolev inequalities that hold for any sufficiently smooth function w with compact support in either R^2 or R^3:

$$\text{(A.4)} \qquad \int \int_{-\infty}^{\infty} w^4 dA \leq \frac{1}{2} \Big(\int \int_{-\infty}^{\infty} w^2 \, dA \Big) \Big(\int \int_{-\infty}^{\infty} w_{,\alpha} w_{,\alpha} \, dA \Big),$$

$$\text{(A.5)} \qquad \int \int \int_{-\infty}^{\infty} w^6 dx \leq \bar{\Omega} \Big(\int \int \int_{-\infty}^{\infty} w_{,j} w_{,j} \, dx \Big)^3.$$

The best constant in (A.5) has been computed to have the value $\bar{\Omega} = (1/27)(2/\pi)^4$. (A.1)–(A.3) are recorded in [1] and (A.4)–(A.5) are listed in [2].

In the following, we also record some properties for the two auxiliary functions which we used in section 3. These functions are exactly the same as ones used in [1] and the derivation of their properties follows from ideas in [1] (see [1, Lemmas 1-3 and 5-7]).

Let R_z and D_z be the notations introduced in section 2. Suppose that φ and Ψ are the solutions of problems (P1) and (P2), respectively.

$$\text{(P1)} \qquad \begin{cases} \Delta\varphi = f \quad \text{in } R_z, \\ \dfrac{\partial\varphi}{\partial\mathbf{n}} = 0 \quad \text{on } \partial R_z, \\ \displaystyle\int_{R_z} f \, dx = 0, \quad f \to 0 \quad \text{as } x_3 \to \infty. \end{cases}$$

$$\text{(P2)} \qquad \begin{cases} \Delta\Psi = 0 \quad \text{in} \quad R_z, \\ \dfrac{\partial\Psi}{\partial\mathbf{n}} = 0 \quad \text{on } \partial D_\xi, \quad \xi \geq z > 0, \\ \dfrac{\partial\Psi}{\partial\mathbf{n}} = g \quad \text{in } D_z \quad \text{with } \displaystyle\int_{D_z} g \, dA = 0. \end{cases}$$

We have the following estimates for the functions φ and Ψ, respectively:

$$\text{(A.6)} \qquad \int_{R_z} \varphi_{,i} \varphi_{,i} \, dx \leq \frac{1}{\mu_2} \int_{R_z} f^2 \, dx,$$

$$\text{(A.7)} \qquad \int_{D_z} \varphi_{,\alpha} \varphi_{,\alpha} \, dx \leq \frac{2}{\sqrt{\mu_2}} \int_{R_z} f^2 \, dx.$$

If, in addition, D_z is star-shaped with respect to a point (origin) in D_z, then

$$\int_z^\infty \int_{\partial D_\xi} |\text{grad}_s \varphi| \, ds d\xi \le \frac{2}{h_0} \left[\frac{1}{\mu_2} + \frac{d^2}{4} \right] \int_{R_z} f^2 \, dx, \tag{A.8}$$

where d is the diameter of D_0 and $h_0 = \min x_\alpha n_\alpha$ on ∂D_ξ. The notation $\text{grad}_s \varphi$ denotes the tangential component of gradient of φ.

Similarly,

$$\int_{D_z} \Psi_{,\alpha} \Psi_{,\alpha} \, dA = \int_{D_z} g^2 \, dA, \tag{A.9}$$

$$\int_{R_z} \Psi_{,i} \Psi_{,i} \, dx \le \frac{1}{\sqrt{\mu_2}} \int_{D_z} g^2 \, dA, \tag{A.10}$$

and if D_z is star-shaped with a point (origin) in D_z, then

$$\int_z^\infty \int_{\partial D_\xi} |\text{grad}_s \Psi| \, ds d\xi \le \frac{2}{h_0} \left[d + \frac{1}{\sqrt{\mu_2}} \right] \int_{D_z} g^2 \, dAr. \tag{A.11}$$

Acknowledgment. The authors would like to thank one of the referees for suggestions leading to the proof of inequality (4.23) which permitted the removal of unnecessary and unphysical restrictions in the original version of this paper.

REFERENCES

[1] K. A. Ames, L. E. Payne, and W. Schaefer, *Spatial decay estimates in time-dependent Stokes flow*, SIAM J. Math. Anal., 24 (1993) pp. 1395–1413.

[2] K. A. Ames and L. E. Payne, *Decay estimates in steady pipe flow*, SIAM J. Math. Anal., 20 (1989), pp. 789–815.

[3] J. N. Flavin, R. J. Knops, and L. E. Payne, *Asymptotic behavior of solutions to semi-linear elliptic equations on the half cylinder*, J. Appl. Math. Phys. (ZAMP), 43 (1992), pp. 405–421.

[4] C. O. Horgan, *Recent developments concerning Saint-Venant's principle: An update*, Appl. Mech. Rev., 42 (1989), pp. 295–303.

[5] C. O. Horgan and J. K. Knowles, *Recent developments concerning Saint-Venant's principle*, Adv. Appl. Mech., 23 (1983), pp. 179–269.

[6] C. O. Horgan and L. E. Payne, *Decay estimates for second order quasilinear partial differential equations*, Adv. Appl. Math., 5 (1984), pp. 309–322.

[7] C. O. Horgan and L. E. Payne, *Decay estimates for a class of nonlinear boundary value problems in two dimensions*, SIAM J. Math. Anal., 20 (1989), pp. 782–788.

[8] C. O. Horgan and L. E. Payne, *On inequalities of Korn, Friedrichs and Babuška-Aziz*, Arch. Rational Mech. Anal., 82 (1983), pp. 165–179.

[9] C. O. Horgan, L. E. Payne, and L. T. Wheeler, *Spatial decay estimates in transient heat conduction*, Quart. Appl. Math., 42, (1984), pp. 119–127.

[10] C. O. Horgan and L. T. Wheeler, *Spatial decay estimates for the Navier–Stokes equations with application to the problem of entry flow*, SIAM J. Math. Anal., 35 (1978), pp. 97–116.

[11] C. T. Hsu and P. Cheng, *Thermal dispersion in a porous medium*, Internat. J. Heat Mass Transfer, 33 (1990), pp. 1587–1597.

[12] C. Lin and L. E. Payne, *Phragmen–Lindelőf type results for second order quasilinear parabolic equations in R^2*, J. Appl. Math. Phys. (ZAMP), 45 (1994), pp. 294–311.

[13] D. A. Nield, *The limitations of the Brinkman–Forchheimer equation in modeling flow in a saturated porous medium and at an interface*, Internat. J. Heat Fluid Flow, 12 (1991).

[14] L. E. Payne, *Uniqueness criteria for steady state solutions of the Navier–Stokes equations*, in Proc. Simpos. Internaz. Appl. Anal. Fis. Mat. (Cagliari-Sassari, 1964), Edizioni Cremonese, Rome, 1965, pp. 130–153.

[15] V. Prasad and N. Kladias, *Non-Darcy natural convection in saturated porous media*, in Convective Heat and Mass Transfer in Porous Media, S. Kakaç, ed., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991, pp. 79–122.

[16] Y. Qin and P. N. Kaloni, *Spatial decay estimates for plane flow in Brinkman–Forcheimer model*, Quart. Appl. Math., to appear.

[17] K. Vafai and C. L. Tien, *Boundary and inertial effects on flow and heat transfer in porous media*, Internat. J. Heat Mass Transfer, 24 (1981), pp. 195–203.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 831–851, July 1997

SELF-SIMILAR SOLUTIONS FOR A MODIFIED BROADWELL MODEL AND ITS FLUID-DYNAMIC LIMITS[*]

HAITAO FAN[†]

Abstract. The existence of self-similar solutions of the Riemann problem for a modified Broadwell model is established. Regularity estimates at the singular points of the problem are obtained. The passing of the fluid-dynamic limit is justified, which yields the Riemann problem for a system of conservation laws.

Key words. Broadwell model, Riemann problem, fluid-dynamic limits, kinetic theory

AMS subject classifications. Primary 82C40, 82C; Secondary 76P05

PII. S0036141095256412

1. Introduction.

In this paper, we study the Riemann problem of the Broadwell model and its fluid-dynamic limits. The Broadwell model, proposed by Broadwell [B], is a system of equations

$$\begin{aligned}\frac{\partial f_1}{\partial t} + \frac{\partial f_1}{\partial x} &= \frac{1}{\epsilon}(f_3^2 - f_1 f_2),\\ \frac{\partial f_2}{\partial t} - \frac{\partial f_2}{\partial x} &= \frac{1}{\epsilon}(f_3^2 - f_1 f_2),\\ \frac{\partial f_3}{\partial t} &= \frac{1}{2\epsilon}(f_1 f_2 - f_3^2)\end{aligned} \tag{1.1}$$

that provides a simple statistical description of a gas of interacting particles. Here the functions f_1 and f_2 are the densities of particles moving in positive and negative x-directions, respectively, and f_3 is the density of particles moving in each of the positive or negative of y- or z-directions. The mean free path ϵ is the measure of average distance between successive collisions. An important feature of this system of equations is its asymptotic equivalence in small mean free path ϵ to the Euler of compressible fluid dynamics

$$\begin{aligned}\rho_t + (\rho u)_x &= 0,\\ (\rho u)_t + (\rho g(u))_x &= 0,\end{aligned} \tag{1.2}$$

where

$$\begin{aligned}\rho &= (f_1 + f_2 + 4(f_1 f_2)^{1/2}),\\ m = \rho u &= f_1 - f_2,\\ g(u) &= \frac{1}{3}[2(1+3u^2)^{1/2} - 1].\end{aligned} \tag{1.3}$$

[*] Received by the editors January 20, 1995; accepted for publication (in revised form) March 11, 1996. This research was supported by an NSF Fellowship under grant DMS-9306064.
http://www.siam.org/journals/sima/28-4/25641.html

[†] Department of Mathematics, Georgetown University, Washington, DC 20057 (fan@gumath1.math.georgetown.edu).

The justification of passing the fluid-dynamic limit in Boltzmann equation or some models of Boltzmann equation has been studied by several authors. The reader is referred to Cercignani [Ce] for a survey of the literature of the Boltzmann equation and to Platkowski and Illner [PI] for results on discrete models of kinetic theory. Recently, Bardos, Golse, and Levermore [BGL] proved the validity of the fluid-dynamic limit of the Boltzmann equation to the incompressible Navier–Stokes equations under some hypothesis. For the Broadwell model, Caflisch and Papanicolaou [CP] and Caflisch [Ca] showed that a given smooth solution of the limit equation can be approximated by a solution of the Broadwell model when ϵ is small. For solutions with shocks of the Broadwell model, there are studies on the stability in time for traveling waves [CL], [KM] and rarefaction waves [Ma]. Xin [X] proved that a given piecewise smooth solution with noninteracting shocks of the limit fluid equations can be approximated by solutions of the Broadwell model as $\epsilon \to 0+$. Recently, Liu and Xin [LX] studied the boundary-layer problems for the Broadwell model and revealed some interesting phenomena. They found that there exist boundary layers in the Broadwell model due to purely kinetic effects that cannot be detected by Chapman–Enskog expansion on the viscous level. They classified the boundary layers as compressive and expansive and showed that expansive boundary layers are stable while compressive boundary layers are stable before they leave the boundary. They also obtained the optimal rate of convergence in the L^∞-norm of kinetic solutions to fluid-dynamic solution in terms of ϵ if the interior fluid flow is smooth. Many of above-mentioned works belong to the approximation program, meaning that an admissible solution of the limit equation, which is a system of conservation laws, is used to construct solutions of the Broadwell model and is intended as a method to solve the Broadwell model. Another approach is to construct solutions of the limit conservation laws as the limit of solutions of the Broadwell model. Recently, Slemrod and Tzavaras [ST], [T] studied the self-similar fluid-dynamic limits of a modified Riemann problem of (1.1) with Maxwellian Riemann data. Work done by Chen and Liu [ChL] and Chen et al. [CLL] on relaxation also sheds light on this subject.

In an attempt to gain insight in the latter direction, Slemrod and Tzavaras [ST] studied the self-similar dynamic-limit approach. For Maxwellian Riemann data

$$f(x,0) = \begin{cases} f_+, & x > 0, \\ f_-, & x < 0, \end{cases} \tag{1.4a}$$

$$Q(f_+) = 0, \qquad Q(f_-) = 0, \quad f_{1\pm}, f_{2\pm}, f_{3\pm} > 0, \tag{1.4b}$$

where

$$Q(f) = f_3^2 - f_1 f_2, \tag{1.4c}$$

the solutions of the limit equation (1.2) are expected to be self-similar functions of $\xi = x/t$. Motivated by this reasoning, they considered the modified Broadwell system

$$\begin{aligned} \frac{\partial f_1}{\partial t} + \frac{\partial f_1}{\partial x} &= \frac{1}{\epsilon t}(f_3^2 - f_1 f_2), \\ \frac{\partial f_2}{\partial t} - \frac{\partial f_2}{\partial x} &= \frac{1}{\epsilon t}(f_3^2 - f_1 f_2), \\ \frac{\partial f_3}{\partial t} &= \frac{1}{2\epsilon t}(f_1 f_2 - f_3^2). \end{aligned} \tag{1.5}$$

By making the ansatz $f(x,t) = f(x/t)$ in (1.4) and (1.5), the Riemann problem (1.4)–(1.5) becomes a singular boundary-value problem:

$$\begin{aligned} (1-\xi)f_1' &= \frac{Q(f)}{\epsilon}, \\ -(1+\xi)f_2' &= \frac{Q(f)}{\epsilon}, \\ \xi f_3' &= \frac{Q(f)}{2\epsilon}, \\ f(-1) = f_-, &\qquad f(+1) = f_+, \\ Q(f_+) = 0, &\qquad Q(f_-) = 0, \quad f_{1\pm}, f_{2\pm}, f_{3\pm} > 0, \end{aligned} \tag{1.6}$$

for $\xi \in [-1,1]$. They proved that the total variations of solutions of (1.6) are bounded uniformly in ϵ, and hence there is a sequence $\{f^{\epsilon_n}\}$, $\epsilon_n \to 0+$, such that $f^{\epsilon_n} \to f$ almost everywhere, where f is a weak solution of the Riemann problem (1.2), (1.4). However, the existence of solutions of (1.6) was proved only in the case where $f_{1-} < f_{1+}$, $f_{2-} > f_{2+}$, and $f_{3-} = f_{3+}$. Under these assumptions on the initial data, the limit solution will be continuous and hence precludes the case where shocks are present.

In this paper, we shall prove the existence of a positive continuous solution of (1.6) and hence (1.4)–(1.5) with no restrictions attached. We also prove some regularity estimates for solutions of (1.6). The precise statement of these results is as follows.

Theorem 1.1. *There exists a positive continuous solution of* (1.6). *Moreover, any positive solution $f(\xi)$ of* (1.6) *satisfies the following estimates:*

$$f_1(\xi) - f_{1+} = O(1)(1-\xi)^{\frac{f_{2+}}{\epsilon}} \quad \textit{for } \xi \textit{near} \xi = 1, \tag{1.7a}$$

$$f_2(\xi) - f_{2-} = O(1)(\xi+1)^{\frac{f_{1-}}{\epsilon}} \quad \textit{for} \xi \textit{ near } \xi = -1, \tag{1.7b}$$

and

$$f_3(\xi) - f_3(0) = O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}} \quad \textit{for } \xi \textit{ near } \xi = 0, \tag{1.7c}$$

$$Q(f(\xi)) = O(1)(\xi+1)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}}. \tag{1.7d}$$

Combining Theorem 1.1 with the results from [ST], we prove the following.

Corollary 1.2. *For any Maxwellian Riemann data* (1.4), *there is a solution of* (1.4)–(1.5) $f^\epsilon(x/t)$. *Further, there is a sequence of solutions of* (1.4)–(1.5), $\{f^{\epsilon_n}\}$, $\epsilon_n \to 0+$, *such that f^{ϵ_n} converges almost everywhere to a weak solution of the limit equations* (1.2) *and* (1.4).

Remark. It is interesting to see that for system (1.1), the discontinuity at $x = 0$ at initial time will propagate along $x/t = \pm 1$ and $x = 0$, which are characteristics of (1.1), while for the modified system (1.5), there is no discontinuity at these locations. In fact, the discontinuities of (1.1) at $x/t = \pm 1$ and $x = 0$ are not intrinsic in the sense that in the limit $\epsilon \to 0+$, there is no shock at $x/t = \pm 1$ due to the stability condition, and there may not be a shock at $x = 0$ in the limit system (1.2).

We organize this paper as follows. In section 2, we recall some results from [ST] and state the main results of this paper. In section 3, we prove that (1.6) has a continuous positive solution. The method we use is a kind of shooting argument. Finally, we prove that for some sequence $\{\epsilon_n\}$, $\epsilon_n \to 0+$ as $n \to \infty$, the solutions of (1.6), f^{ϵ_n}, converge almost everywhere to a weak solution of (1.2), (1.4).

2. Preliminaries.

We see that there are three singular points, $\xi = \pm 1, 0$, in the boundary-value problem (1.6). At $\xi = \pm 1$, two components of f are continuous. Thus the other component must also be continuous due to the boundary condition. Therefore, in what follows, we define the solutions of (1.6) to be weak solutions of (1.6) which are continuous on $[-1, 0) \cup (0, 1]$.

We recall some results on (1.6) obtained in [ST].

Lemma 2.1. *Let $f = (f_1, f_2, f_3)$ be a continuous solution of* (1.6). *Then*

(i) $Q(f(\xi))$ *does not change sign on the intervals* $(-1, 0)$ *or* $(0, 1)$*,*

(ii) $Q(f(-1)) = Q(f(0)) = Q(f(+1)) = 0$*, and*

(iii) f_1*,* f_2*, and* f_3 *are uniformly bounded from above and below by positive constants and of uniformly bounded total variation on* $[-1, 1]$. *These bounds are independent of* $\epsilon > 0$.

In fact, assertion (i) of Lemma 2.1 does not require the continuity of f. For our later use, we revise it as follows.

Lemma 2.2. *Let f be a solution of* (1.6). *Then $Q(f(\xi))$ does not change sign on the interval* $(-1, 0)$ *and* $(0, 1)$. *Furthermore, each component of f is monotone on each interval* $(-1, 0)$ *and* $(0, 1)$.

Proof. A straightforward calculation based on (3.1) shows that

$$\frac{dQ}{d\xi} = \frac{1}{\epsilon}\left(\frac{f_1(\xi)}{\xi + 1} + \frac{f_2(\xi)}{\xi - 1} + \frac{f_3(\xi)}{\xi}\right) Q(\xi), \tag{2.1}$$

from which the assertion follows. □

3. Existence of solutions of (1.6).

In this section, when we refer to solutions of a system of ordinary differential equations, we always mean continuous solutions unless otherwise indicated. We intend to prove the existence of (1.6) by a kind of shooting argument. For this purpose, we consider the trajectories of

$$\begin{aligned} (1 - \xi) f_1' &= \frac{Q(f)}{\epsilon}, \\ -(\xi + 1) f_2' &= \frac{Q(f)}{\epsilon}, \\ \xi f_3' &= \frac{Q(f)}{2\epsilon}, \quad \xi \in (-1, 1), \end{aligned} \tag{3.1}$$

issued from

$$(f_{1-}, f_{2-}, f_{3-}) \quad \text{at } \xi = -1 \tag{3.2}$$

and

$$(f_{1+}, f_{2+}, f_{3+}) \quad \text{at } \xi = +1, \tag{3.3}$$

respectively, where $f_{3\pm} = (f_{1\pm} f_{2\pm})^{1/2}$. It is clear that (3.1) has three singular points, $\xi = -1, 0, +1$. We first need some regularity results of trajectories of (3.1) near these points.

Lemma 3.1.

(i) *Let f be a bounded solution of* (3.1) *and* (3.2) *on* $[-1, 0)$. *Then*

(3.4a)

$$Q = C(1 + \xi)^{\frac{f_{1-}}{\epsilon}} (1 - \xi)^{\frac{f_{2+}}{\epsilon}} |\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}} \times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta) - f_{1-}}{\zeta + 1} + \frac{f_2(\zeta) - f_{2+}}{\zeta - 1} + \frac{f_3(\zeta) - \min(f_{3-}, f_3(0-))}{\zeta}\right) d\zeta\right].$$

In fact, the numbers f_1*,* f_{2+}*, and* $\min(f_{3-}, f_3(0-))$ *in* (3.4a) *can be replaced by any other numbers.*

(ii) *Let* f *be a bounded solution of* (3.1) *and* (3.2) *on* $(0, 1]$*. Then*

$$Q = C(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\max(f_{3+}, f_3(0+))}{\epsilon}} \times \exp\left[\frac{1}{\epsilon}\int_1^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1}+\frac{f_2(\zeta)-f_{2+}}{\zeta-1}+\frac{f_3(\zeta)-\max(f_{3+}, f_3(0+))}{\zeta}\right)d\zeta\right]. \tag{3.4b}$$

In fact, the numbers f_{1-}*,* f_{2+}*, and* $\max(f_{3+}, f_3(0+))$ *in* (3.4b) *can be replaced by any other numbers.*

(iii) *Let* f *be a bounded solution of* (3.1) *and* (3.2) *on* $[-1, 0)$*. Then*

$$Q \le O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}}. \tag{3.5a}$$

(iv) *Let* f *be a bounded solution of* (3.1) *and* (3.3) *on* $(0, 1]$*. Then*

$$Q \le O(1)|\xi|^{\frac{\max(f_{3+}, f_3(0+))}{\epsilon}}. \tag{3.5b}$$

Proof. (i) Since f_3 is bounded and monotone on each of the intervals $[-1, 0)$ and $(0, 1]$, the one-sided limits $f_3(0\pm)$ are well defined and finite. Equation (3.4a) can be obtained from (2.1) as follows:

$$\begin{aligned} Q(f(\epsilon)) &= Q\left(f\left(-\frac{1}{2}\right)\right) \times \exp\left[\frac{1}{\epsilon}\int_{-\frac{1}{2}}^{\xi}\left(\frac{f_1(\zeta)}{\zeta+1}+\frac{f_2(\zeta)}{\zeta-1}+\frac{f_3(\zeta)}{\zeta}\right)d\zeta\right] \\ &= Q\left(f\left(-\frac{1}{2}\right)\right)(\xi+1)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0-)}{\epsilon}} \\ &\quad \times \exp\left[\frac{1}{\epsilon}\int_{\frac{-1}{2}}^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1}+\frac{f_2(\zeta)-f_{2+}}{\zeta-1}+\frac{f_3(\zeta)-\min(f_{3-}, f_3(0-))}{\zeta}\right)d\zeta\right] \\ &= C(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}} \\ &\quad \times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1}+\frac{f_2(\zeta)-f_{2+}}{\zeta-1}+\frac{f_3(\zeta)-\min(f_{3-}, f_3(0-))}{\zeta}\right)d\zeta\right]. \end{aligned} \tag{3.6}$$

The proof of (ii) is similar.

(iii) If $Q \ge 0$ on $[-1, 0)$, then f_2 and f_3 are decreasing and

$$f_1' = \frac{Q}{\epsilon(1-\xi)} \le \frac{f_3^2}{\epsilon(1-\xi)} \le \frac{f_{3-}^2}{\epsilon(1-\xi)},$$

and hence the integral in (3.4a) is either negative or bounded. Thus we have the estimate

$$Q \le O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}}.$$

If $Q < 0$ on $[-1, 0)$, then f_1 is decreasing and f_2 is increasing; hence $f_1(\xi) < f_{1-}$ and $f_2(\xi > f_{2-}$ for $\xi \in [-1, 0)$. A variation of (3.4a) states that

$$Q = C(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}} \times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1}+\frac{f_2(\zeta)-f_{2-}}{\zeta-1}+\frac{f_3(\zeta)-\min(f_{3-}, f_3(0-)}{\zeta}\right)d\zeta\right].$$

Since every term of the integrand is negative, we also have

$$Q = O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}}.$$

We can prove (iv) similarly. □

Now we need some existence results for the trajectories of (3.1) and (3.2) and those of (3.1) and (3.3). For definiteness, we consider (3.1) and (3.2). The other part can be handled similarly. In view of Lemma 3.1, we let

$$P(\xi) := \frac{Q(f(\xi))}{(1+\xi)^{\frac{f_{1-}}{\epsilon}}}. \tag{3.7}$$

Under this transformation, the initial-value problem (3.1)–(3.2) becomes a regular problem:

$$\begin{aligned} f_1' &= \frac{(1+\xi)^{\frac{f_{1-}}{\epsilon}} P}{\epsilon(1-\xi)}, \\ f_2' &= \frac{-(1+\xi)^{\frac{f_{1-}}{\epsilon}-1} P}{\epsilon}, \\ P' &= \frac{P(\xi)}{\epsilon}\left(\frac{1}{1+\xi}\int_{-1}^{\xi} \frac{(1+s)^{\frac{f_{1-}}{\epsilon}}}{(1-s)\epsilon} P(s)ds \right. \\ &\quad \left. + \frac{f_2}{\xi-1} + \frac{f_{3-} + \int_{-1}^{\xi} \frac{1}{2\epsilon s}(1+s)^{\frac{f_{1-}}{\epsilon}} P(s)ds}{\xi}\right) \end{aligned} \tag{3.8}$$

with initial conditions

$$f_1(-1) = f_{1-}, f_2(-1) = f_{2-}, P(-1) = P_-. \tag{3.9}$$

LEMMA 3.2. *The system of equations* (3.8)–(3.9) *is equivalent to* (3.1)–(3.2).

Proof. It is clear that (3.1) implies (3.8). To see that (3.8) implies (3.1), we let

$$\begin{aligned} Q &= (1+\xi)^{\frac{f_{1-}}{\epsilon}} P(\xi), \\ f_3 &:= f_{3-} + \int_{-1}^{\xi} \frac{Q}{\epsilon s} ds. \end{aligned} \tag{3.10}$$

A straightforward calculation based on (3.8) yields (3.1) and

$$\begin{aligned} \frac{dQ}{d\xi} &= \frac{1}{\epsilon}\left(\frac{f_1}{\xi+1} + \frac{f_2}{\xi-1} + \frac{f_3}{\xi}\right)Q \\ &= -f_1 f_2' - f_2 f_1' + (f_3^2)' \\ &= \frac{d}{d\xi}(f_3^2 - f_1 f_2), \end{aligned}$$

which implies

$$\begin{aligned} Q &= f_2^2 - f_1 f_2 + f_3^2(-1) - f_1(-1) f_2(-1) \\ &= f_3^2 - f_1 f_2. \qquad \square \end{aligned} \tag{3.11}$$

The following lemma indicates how $f(\xi, P_-)$ can be extended to $[-1, 0]$.

LEMMA 3.3. *For any $f_{1-} > 0$, $f_{2-} > 0$, and $P_- = (Q(f(\xi))/(1+\xi)^{\frac{f_{1-}}{\epsilon}})\big|_{\xi=-1}$, problem* (3.8)–(3.9) *has a unique solution on* $[-1, \xi_0)$ *for some* $\xi_0 \in (-1, 0]$.

Proof. System (3.8)–(3.9) is an initial-value problem for systems of regular differential integral equations to which the standard contraction-mapping argument applies, and hence we have the assertion. □

For convenience, we shall denote the trajectory of (3.1)–(3.2) that satisfies (3.18) by

$$f(\xi; f_{1-}, f_{2-}, P_-)$$

or by $f(\xi; P_-)$ when no confusion will arise.

LEMMA 3.4. *Let $f_{1-} > 0$ and $f_{2-} > 0$.*

(i) *If $P_- \le 0$, then problem* (3.8)–(3.9) *has a unique solution on* $[-1, 0]$.

(ii) *For $P_- > 0$, the solution of* (3.8)–(3.9) *exists on* $[-1, 0]$ *if and only if* $f_2(\xi) > 0$ *on* $[-1, 0) \cap$ (*domain of* f).

(iii) *If the solution of* (3.1)–(3.2) *exists on* $[-1, 0]$, *then f is positive on* $[-1, 0)$, *i.e., all of the components of f are positive on* $[-1, 0)$.

Proof. From the structure of (3.1), we can see that if $Q(f(\xi))$ is bounded on $[-1, 0]$ in supremum norm, then $f(\xi)$ is bounded in $[-1, \xi_1]$ for any $\xi_1 \in (-1, 0)$. Then by the standard argument of continuation of contraction mapping, the existence and uniqueness result of Lemma 3.3 can be extended to $[-1, 0)$.

(i) In the case where $P_- \le 0$, we have similarly to (3.4a) that

$$Q(f(\xi)) = P_-(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{f_{3-}}{\epsilon}} \times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1} + \frac{f_2(\zeta)-f_{2-}}{\zeta-1} + \frac{f_3(\zeta)-f_{3-}}{\zeta}\right)d\zeta\right] < 0. \tag{3.12}$$

Then equation (3.1) implies that

$$\begin{aligned} f_1(\zeta) &< f_{1-} = f_1(-1),\\ f_2(\zeta) &> f_{2-} = f_2(-1),\\ f_3(\zeta) &> f_{3-} \end{aligned} \tag{3.13}$$

for $\zeta \in (-1, 0]$ and therefore

$$|Q(f(\xi))| \le |P_-|(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{f_{3-}}{\epsilon}}.$$

Hence f is bounded in $[-1, 0]$, and thus the solution of (3.8)–(3.9) exists on $(-1, 0)$. Further, because of (3.13), the existence can be extended to $[-1, 0]$.

(ii) In the case where $P_- > 0$, we can derive from (3.12) that $Q(f(\xi)) > 0$ on $[-1, 0]$. Then (3.1) shows that

$$\begin{aligned} f_1(\xi) &> f_{1-} > 0,\\ f_2(\xi) &< f_{2-},\\ f_3(\xi) &< f_{3-} \end{aligned}$$

for $\xi \in (-1, 0) \cap$ (domain of existence of f). If $f_2(\xi) > 0$ for $\xi \in (-1, 0)$, then

$$f_{3-} > f_3(\xi) > 0, \quad \xi \in (-1, 0),$$

because otherwise equation $(3.1)_3$ shows that at ξ_0, the infimum of points at which $f_3(\xi)=0$,

$$f_3'(\xi_0)=\frac{-f_1(\xi_0)f_2(\xi_0)}{\epsilon\xi_0}>0,$$

which cannot be true at ξ_0. Thus we have

$$0<Q(f(\xi))<f_3^2(\xi)<f_{3-}^2. \tag{3.14}$$

Conversely, if $f_2(\xi_0)<0$ for some $\xi_0\in(-1,0)$, then

$$f_2'(\xi)=\frac{Q(f)}{\xi}<\frac{-f_1(\xi_0)f_2(\xi_0)}{\xi}<0,$$

where we have used the monotonicity of f_1 and f_2 on $(-1,0)$. It can be seen that $f_2(\xi)\to-\infty$ and hence $Q(f(\xi))\to\infty$ as ξ approaches $0-$. Then the assertion in (ii) follows.

(iii) We first claim that $f_1(\xi)>0$ on $[-1,0)$. Indeed, otherwise, there would be a point $\xi_0\in[-1,0)$ such that $f_1(\xi_0)=0$ and $f_1(\xi)>0$ for $\xi\in[-1,\xi_0)$. Then we have $f_1'(\xi_0)\le 0$. On the other hand, however, equation $(3.1)_1$ implies that $0\le(1-\xi_0)f_1'(\xi_0)=Q(f(\xi_0))=f_3^2(\xi_0)\ge 0$. From (3.4a), we see that $Q(f(\xi)$ remains positive or negative or 0 on $(-1,0)$, which in our case says that $Q\equiv 0$ on $(-1,0)$. This implies that $f_1\equiv f_{1-}$ on $(-1,0)$, which is a contradiction.

When $P_-\le 0$, our assertion holds in view of (3.13).

When $P_->0$, assertion (ii) says that $f_2(\xi)>0$ on $[-1,0)$. Further, the proof of (ii) indicates that if $f_2(\xi)>0$ on $[-1,0)$, then so does f_3. This completes the proof of (iii). □

Theorem 3.5. *For any $f_{1-}>0$ and $f_{2-}>0$, there is P_0, $0<P_0<\infty$, such that problem* (3.8)–(3.9) *has a unique solution on $[-1,0]$ for $P_-\in(-\infty,P_0]$ and*

$$f_2(0;f_{1-},f_{2-},P_0)=0\,. \tag{3.15}$$

Furthermore, the solution $f(\xi,P_-)$ is positive on $[-1,0)$ for all $P_-\in(-\infty,P_0]$.

Proof. From Lemma 3.4(i), we see that problem (3.8)–(3.9) has a unique solution on $[-1,0]$ for all $P_-\le 0$.

Suppose (3.8)–(3.9) has a solution for some $\bar{P}_->0$. We claim that (3.8)–(3.9) has a solution for all $P_-\in(-\infty,\bar{P}_-]$. To this end, without loss of generality, we consider only $\bar{P}_->P_->0$. Let

$$\begin{aligned}\bar{f}(\xi)&:=f(\xi;f_{1-},f_{2-},\bar{P}_-),\\ f(\xi)&:=f(\xi;f_{1-},f_{2-},P_-)\,.\end{aligned}$$

We claim that

$$Q(\bar{f}(\xi))>Q(f(\xi)) \tag{3.16}$$

for all $\xi\in(-1,0)$ if $f(\xi)$ exists. To see this, we recall that

$$\bar{P}(-1)=\frac{Q(f(\xi))}{(1+\xi)^{\frac{f_{1-}}{\epsilon}}}\bigg|_{\xi=-1}=\bar{P}_->P_-=P(-1)=\frac{Q(f(\xi))}{(1+\xi)^{\frac{f_{1-}}{\epsilon}}}\bigg|_{\xi=-1}.$$

Then there is a $\xi_0 \in (-1, 0]$ such that for $\xi \in (-1, \xi_0)$, the solutions $\bar{P}(\xi)$ and $P(\xi)$ to (3.8)–(3.9), with $\bar{P}(-1) = \bar{P}_-$ and $P(-1) = P_-$, respectively, satisfy

$$\bar{P}(\xi) > P(\xi)$$

and hence

$$Q(\bar{f}(\xi)) > Q(f(\xi)).$$

Without loss of generality, we can assume ξ_0 to be the maximum of ξ_0 in the above statement. Then

$$Q(\bar{f}(\xi_0)) = Q(f(\xi_0)) \tag{3.17}$$

and

$$Q(\bar{f}(\xi)) > Q(f(\xi)) \tag{3.18}$$

for $\xi \in (-1, \xi_0)$. From (3.1), we see that

$$\begin{aligned} &\bar{f}_1(\xi) > f(\xi),\\ &\bar{f}_2(\xi) < f_2(\xi),\\ &\bar{f}_3(\xi) < f_3(\xi) \end{aligned} \tag{3.19}$$

for $\xi \in (-1, \xi_0)$. Using the same technique used to derive (3.4), we obtain

$$\begin{aligned} 0 &= Q(\bar{f}(\xi_0)) - Q(f(\xi_0))\\ &= \bar{P}_-(1+\xi)^{\frac{f_{1-}}{\epsilon}} \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi_0}\left(\frac{\bar{f}_1(\zeta) - f_{1-}}{\zeta+1} + \frac{\bar{f}_2(\zeta)}{\zeta-1} + \frac{\bar{f}_3(\zeta)}{\zeta}\right)d\zeta\right]\\ &\quad - P_-(1+\xi)^{\frac{f_{1-}}{\epsilon}} \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta) - f_{1-}}{\zeta+1} + \frac{f_2(\zeta)}{\zeta-1} + \frac{f_3(\zeta)}{\zeta}\right)d\zeta\right]\\ &= Q(f(\xi_0))\left\{\frac{\bar{P}_-}{P_-}\exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{\bar{f}_1(\zeta) - f_1(\zeta)}{\zeta+1} + \frac{\bar{f}_2(\zeta) - f_2(\zeta)}{\zeta-1}\right.\right.\right.\\ &\qquad\qquad \left.\left.\left. + \frac{\bar{f}_3(\zeta) - f_3(\zeta)}{\zeta}\right)d\zeta\right] - 1\right\} > 0, \end{aligned}$$

where in the last inequality we used (3.19), which is a contradiction. This proves that if $\bar{P}_- > P_- > 0$,

$$0 < Q(f(\xi)) < Q(\bar{f}(\xi))\,. \tag{3.20}$$

Then by (3.1) and the assumption that $\bar{f}(\xi)$ is a solution of (3.1)–(3.2) on $[-1, 0]$, we have

$$0 \le \bar{f}_2(\xi) < f_2(\xi),$$

and hence $f(\xi, P_-)$ exists in view of Lemma 3.4(ii).

Now we denote

$$\begin{aligned} P_0 := \sup\{\bar{P}_- \ge 0;\ &(3.8)\text{–}(3.9) \text{ has a solution on}\\ &[-1, 0] \text{ with } P_- = \bar{P}_- \text{ on } [-1, 0]\} \ge 0. \end{aligned} \tag{3.21}$$

Then we have proved that (3.8)–(3.9) has a solution on $[-1,0]$ with $P_- \in (-\infty, P_0)$. It remains to prove that $P_0 < +\infty$ and that $f(\xi, f_{1-}, f_{2-}, P_0)$ exists on $[-1,0]$ and $f_2(0,P_0) = 0$. We claim that

$$\lim_{P_- \to P_0} f(\xi; P) \tag{3.22}$$

exists for all $\xi \in [-1,0]$ and is the solution $f(\xi, f_{1-}, f_{2-}, P_0)$. To this end, we observe that $P_0 \geq 0$. If $-1 \leq P \leq 0$, then the proof of Lemma 3.4(i) states that $f(\xi)$ is bounded on $[-1,0]$. If $P_0 > P > 0$, then by (3.19) we have

$$0 \leq Q(f(\xi; P_-)) < f_3^2(\xi; P_-) < f_{3-}^2, \tag{3.23}$$

and hence $f_1(\xi; f_{1-}, f_{2-}, P_-)$ is bounded uniformly in P_-, and so are f_2 and f_3 since

$$\begin{aligned} 0 &< f_2(\xi; P_-) < f_{2-}, \\ 0 &< f_3(\xi; P_-) < f_{3-}. \end{aligned} \tag{3.24}$$

Also, f_1, f_2, and f_3 are monotone in ξ on $[-1,0]$ and hence have total variation bounded independently of $P \in [-1, P_0)$, and hence (3.22) exists. The conclusion that the limit is the solution $f(\xi; f_{1-}, f_{2-}, P_0)$ is obvious from the integral form of (3.8). We claim that $P_0 < +\infty$. Indeed, otherwise, we would have similarly to (3.4) that

$$\begin{aligned} Q(f(\xi; P_-)) &= P_-(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{f_{3-}}{\epsilon}} \\ &\quad\times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta)-f_{1-}}{\zeta+1} + \frac{f_2(\zeta)-f_{2-}}{\zeta-1} + \frac{f_3(\zeta)-f_{3-}}{\zeta}\right)d\zeta\right] \\ &\quad> P_-(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{f_{3-}}{\epsilon}} \to \infty \quad \text{as } P_- \to +\infty \end{aligned}$$

for $\xi \in (-1,0)$, where we used (3.24) when $P_- > 0$. Then from $(3.1)_2$, we have

$$f_2(\xi; P_-) = f_{2-} - \int_{-1}^{\xi} \frac{Q(f(\zeta; P_-))}{\zeta+1} d\zeta \to -\infty \qquad \text{as } P_- \to +\infty \tag{3.25a}$$

for $\xi \in (-1,0)$. On the other hand, we have

$$f_2(\xi, P_0) \geq f_2(0; P_0) = \lim_{P_- \to P_0} f_2(0; P_-) \geq 0 \tag{3.25b}$$

for $\xi \in [-1,0]$. The contradiction between (3.25a) and (3.25b) proves the claim that $P_0 < +\infty$. Finally, we verify that

$$f_2(0, P_0) = 0\,.$$

For contradiction, we assume the contrary, which is, in light of (3.25b), that

$$f_2(0; P_0) > 0. \tag{3.26}$$

Consider $f(\xi; f_{1-}, f_{2-}, P_-)$ for any P_- close to P_0 and $P_- > P_0 \geq 0$. The solution $f(\xi; P_-)$ of (3.8)–(3.9) cannot be defined on $[-1,0]$, for otherwise the definition of P_0, (3.21), would be violated. Then by Lemma 3.4(ii),

$$f_2(\xi_0; f_{1-}, f_{2-}, P_-) = 0 \tag{3.27}$$

for some $\xi_0 = \xi_0(P_-) \in (-1, 0)$. Take a sequence of $\{P_{-,k}\}_{k=1}^{\infty}$ such that $P_{-,k} \to P_0+$ and $\xi_k = \xi_0(P_{-,k}) \to \eta$ as $k \to +\infty$. For any small $\tau > 0$, by the definition of η, $f(\xi; f_{1-}, f_{2-}, P_{-,k})$ exists on $[-1, -\tau + \eta]$. Since the right-hand side of (3.8) is continuous and hence its solution is continuous in initial value (3.9), we have

$$|f(-\tau + \eta, P_0) - f(-\tau + \eta, P_{-,k})| \to 0 \quad \text{as } k \to \infty. \tag{3.28}$$

Since $f_2(\xi)$ is decreasing when $P_- > 0$, we infer from (3.26) that

$$f_2(-\tau + \eta, f_{1-}, f_{2-}, P_0) =: f_0 > 0, \tag*{(3.29),}$$

and hence

$$f_2(-\tau + \eta, P_{-,k}) > \frac{1}{2} f_0 \tag{3.30}$$

for large k. Then the mean-value theorem implies that

$$\begin{aligned} |f_2'(\theta; P_{-,k})| &= \left| \frac{f_2(\xi_0(P_{-,k}), P_{-,k}) - f_2(-\tau + \eta, P_{-,k})}{\xi_0(P_{-,k}) + \tau - \eta} \right| \\ &\geq \frac{|0 - \frac{1}{2} f_0|}{2\tau} = \frac{f_0}{4\tau} \end{aligned} \tag{3.31}$$

for some $\theta \in (-\tau + \eta, \xi_0(P_{-,k})) \subset (-1, 0)$. On the other hand, we also have

$$\begin{aligned} |f_2'(\theta; P_{-,k})| &= \frac{Q(f(\theta; P_{-,k}))}{(\theta + 1)\epsilon} \\ &< \frac{f_3^2(\theta, P_{-,k})}{(\theta + 1)\epsilon} < \frac{f_{3-}^2}{(\theta + 1)\epsilon} < \frac{f_{3-}^2}{(1 - \tau + \eta)\epsilon}, \end{aligned} \tag{3.32}$$

where we have used the fact that $f_2(\xi, P_{-,k}) > 0$ and $0 < f_3(\xi, P_{-,k}) < f_{3-}$ for $\xi \in (-\tau + \eta, \xi_0(P_{-,k}))$. Combining (3.31) and (3.32), we obtain a contradiction:

$$\frac{f_{3-}^2}{(1 - \tau + \eta)\epsilon} > \frac{f_0}{4\tau}, \tag{3.33}$$

where $\tau > 0$ can be arbitrarily small. This contradiction proves (3.15). □

We shall establish the Lipschitz-continuous dependence on P_- of $f(\xi, P_-)$. Since the right-hand side of (3.8) is not Lipschitz continuous at $\xi = -1$ and 0, we have to prove it directly.

LEMMA 3.6. *Suppose* $f(\xi) = f(\xi; f_{1-}, f_{2-}, P_-)$ *and* $\bar{f}(\xi) = f(\xi; f_{1-}, f_{2-}, \bar{P}_-)$ *exist on* $[-1, 0]$ *and* $f_2(0-) > 0$ $\bar{f}_2(0-) > 0$. *Then for all* $\xi \in [-1, 0]$,

$$|f(\xi; f_{1-}, f_{2-}, P_-) - f(\xi; f_{1-}, f_{2-}, \bar{P}_-)| \leq C|P_- - \bar{P}_-| \tag{3.34}$$

for some constant $C > 0$.

Proof. By the definition in (3.7), we have

$$P(\xi) = P_- \exp\left[\frac{1}{\epsilon} \int_{-1}^{\xi} \left(\frac{f_1(\zeta) - f_{1-}}{\zeta + 1} + \frac{f_2(\zeta)}{\zeta - 1} + \frac{f_3(\zeta)}{\zeta} \right) d\zeta \right].$$

Consequently,

$$\bar{P}(\xi) - P(\xi) = \frac{P(\xi)}{P_-}\Big\{\bar{P}_- \exp\Big[\frac{1}{\epsilon}\int_{-1}^{\xi}\Big(\frac{\bar{f}_1(\zeta) - f_1(\zeta)}{\zeta+1} + \frac{\bar{f}_2(\zeta) - f_2(\zeta)}{\zeta-1} + \frac{\bar{f}_3(\zeta) - f_3(\zeta)}{\zeta}\Big)d\zeta\Big] - P_-\Big\},$$

where we assumed without loss of generality that $P_- \neq 0$. This leads to

(3.35)
$$\begin{aligned}\bar{P}(\xi) - P(\xi) &= \frac{P(\xi)}{P_-}(\bar{P}_- - P_-) \\ &\quad + \frac{P(\xi)\bar{P}_-}{P_-}\Big\{\exp\Big[\frac{1}{\epsilon}\int_{-1}^{\xi}\Big(\frac{\bar{f}_1(\zeta) - f_1(\zeta)}{\zeta+1} + \frac{\bar{f}_2(\zeta) - f_2(\zeta)}{\zeta-1} + \frac{\bar{f}_3(\zeta) - f_3(\zeta)}{\zeta}\Big)d\zeta\Big] - 1\Big\} \\ &= \frac{P(\xi)}{P_-}(\bar{P}_- - P_-) \\ &\quad + \frac{P(\xi)\bar{P}_-}{P_-}\Big\{\exp\Big[\frac{1}{\epsilon^2}\int_{-1}^{\xi}\int_{-1}^{\zeta}\Big(\frac{(1+s)^{\frac{f_{1-}}{\epsilon}}}{(\zeta+1)(1-s)} + \frac{(1+s)^{\frac{f_{1-}}{\epsilon}-1}}{(\zeta-1)} \\ &\quad + \frac{(1+s)^{\frac{f_{1-}}{2}}}{s\zeta}\Big)(\bar{P}(s) - P(s))ds d\zeta\Big] - 1\Big\}.\end{aligned}$$

We denote

$$\alpha = \frac{\min(f_{3-}, f_3(0-), \bar{f}_3(0-))}{\epsilon}.$$

By the structure of (3.1), the assumption that $f_2(0-) > 0$ and $\bar{f}_2(0-) > 0$ implies that $f_3(0-) > 0$ and $\bar{f}_3(0-) > 0$ and hence that $\alpha > 0$. Furthermore, from (3.5), we see that $P(\xi)/\xi^\alpha$ is bounded. We divide (3.35) by ξ^α to obtain

(3.36)
$$\begin{aligned}\Big|\frac{\bar{P}(\xi) - P(\xi)}{\xi^\alpha}\Big| &\leq |\bar{P}_- - P_-|\Big|\frac{P(\xi)}{P_-\xi^\alpha}\Big| \\ &\quad + \Big|\frac{P(\xi)\bar{P}_-}{\xi^\alpha P_-}\Big|\Big\{\exp\Big[\frac{1}{\epsilon^2}\int_{-1}^{\xi}d\zeta\int_{-1}^{\zeta}\Big(\frac{\frac{(1+s)^{f_{1-}}}{2}}{(\zeta+1)(1-s)} \\ &\quad + \frac{(1+s)^{\frac{f_{1-}}{\epsilon}} - 1}{1-\zeta} - \frac{(1+s)^{\frac{f_{1-}}{2}}}{\zeta}\Big)s^{\alpha-1}\Big|\frac{\bar{P}(s) - P(s)}{s^\alpha}\Big|ds\Big] - 1\Big\} \\ &\leq |\bar{P}_- - P_-|\Big|\frac{P(\xi)}{\xi^\alpha P_-}\Big| \\ &\quad + \Big|\frac{P(\xi)\bar{P}_-}{\xi^\alpha P_-}\Big|\Big\{\exp\Big[M(\alpha)\int_{-1}^{\xi}\max_{s\in[-1,\zeta]}\Big|\frac{\bar{P}(s) - P(s)}{s^\alpha}\Big|d\zeta\Big] - 1\Big\},\end{aligned}$$

where $M(\alpha)$ is a constant bounded for $\alpha > \delta > 0$. This yields that for $\xi \in [-1, 0]$,

$$\begin{aligned}\max_{s\in[-1,\xi]}\left|\frac{\bar{P}(s)-P(s)}{s^{\alpha}}\right| &\leq |P_{-}-\bar{P}_{-}|\max_{s\in[-1,\xi]}\left|\frac{P(s)}{s^{\alpha}P_{-}}\right| \\ &\quad + \max_{s\in[-1,\xi]}\left|\frac{P(s)\bar{P}_{-}}{s^{\alpha}P_{-}}\right|\exp\left(M\int_{-1}^{\xi}\max_{s\in[-1,\zeta]}\left|\frac{\bar{P}(s)-P(s)}{s^{\alpha}}\right|d\zeta\right) \\ &\quad \times\int_{-1}^{\xi}\max_{s\in[-1,\zeta]}\left|\frac{\bar{P}(s)-P(s)}{s^{\alpha}}\right|d\zeta \\ &\leq A+B\int_{-1}^{\xi}\max_{s\in[-1,\zeta]}\left|\frac{\bar{P}(s)-P(s)}{s^{\alpha}}\right|d\zeta.\end{aligned} \tag{3.37}$$

Applying Gronwall's inequality to the above expression, we obtain

$$\max_{s\in[-1,\xi]}\left|\frac{\bar{P}(s)-P(s)}{s^{\alpha}}\right| \leq |\bar{P}_{-}-P_{-}|A\exp[B(\xi+1)]$$

for $\xi \in (-1, 0]$. Then the assertion follows from (3.8), (3.7), and (3.1). □

COROLLARY 3.7. *Let P_0 be as in Theorem 3.5. Then $f(0, f_{1-}, f_{2-}, P_{-})$ is Lipschitz continuous in P_{-} for $P_{-} \in (-\infty, P_0 - \delta]$ for any $\delta > 0$.*

We see from the last theorem that

$$\left\{(f_1(0; f_{1-}, f_{2-}, P_{-}), f_2(0; f_{1-}, f_{2-}, P_{-})) : P_{-} \in (-\infty, P_0)\right\} =: C_{-}(f_{1-}, f_{2-}) \tag{3.38}$$

is a continuous curve in the first quadrant of the (f_1, f_2)-plane for each fixed (f_{1-}, f_{2-}). Taking $P_{-} = 0$, we see that

$$(f_{1-}, f_{2-}) \in C_{-}(f_{1-}, f_{2-}). \tag{3.39}$$

We study the range of this curve in the following theorem.

THEOREM 3.8. *Let $C_{-}(f_{1-}, f_{2-})$ be defined as in (3.36). Then*

$$C_{-}(f_{1-}, f_{2-}) \subset \left\{(f_1, f_2) \in \mathbb{R}^2 : 0 < f_1 \leq f_1(0; P_0), 0 = f_2(0; P_0) \leq f_2\right\}. \tag{3.40}$$

Furthermore,

$$f_2(0, P_{-}) \to +\infty \text{ as } P_{-} \to -\infty. \tag{3.41}$$

Proof. From Lemma 3.4(ii), we see that $f_2(0; P_{-}) \geq 0$ for $P_{-} \in (-\infty, P_0]$. The structure of (3.1) guarantees

$$f_1(0, P_{-}) > 0.$$

To prove (3.40), it suffices to prove that

$$f_1(0; P_{-}) \leq f_1(0; P_0). \tag{3.42}$$

If $P_{-} \leq 0$, then $Q(f(\xi, P_{-})) \leq 0$ on $(-1, 0)$. Equation (3.1) then implies that

$$f_1(0; P_{-}) \leq f_{1-} = f_1(-1; 0). \tag{3.43}$$

For $P_- \in [0;P_0]$, we recall that $Q(f(\xi,P_-))$ is monotone increasing with respect to $P_- > 0$; see (3.16). Equation (3.1) then states that $f_1(0,P_-)$ is increasing with respect to $P_- > 0$. Thus

$$f_{1-} \le f_1(0;P_-) \le f_1(0;P_0) \tag{3.44}$$

for $P_- \in [0,P_0]$. Inequality (3.42) follows from (3.43) and (3.44).

To prove (3.41), we let

$$P_- < 0$$

and consider the following application of (3.4).

$$\begin{aligned}
f_2(\xi;P_-) - f_{2-} &= \frac{-1}{\epsilon}\int_{-1}^{\xi} \frac{Q(f(\zeta))}{\zeta+1}\,d\zeta \\
&= \frac{-P_-}{\epsilon}\int_{-1}^{\xi}(1+\zeta_1)^{\frac{f_{1-}}{\epsilon}-1}(1-\zeta_1)^{\frac{f_{2-}}{\epsilon}}|\zeta_1|^{\frac{f_{3-}}{\epsilon}} \\
\times \exp\biggl[\frac{1}{\epsilon}\int_{-1}^{\zeta_1}\biggl(&\frac{f_1(\zeta_2)-f_{1-}}{\zeta_2+1} + \frac{f_2(\zeta_2)-f_{2-}}{\zeta_2-1} + \frac{f_3(\zeta_2)-f_{3-}}{\zeta_2}\biggr)d\zeta_2\biggr]d\zeta_1 \\
&= \frac{-P_-}{\epsilon}\int_{-1}^{\xi}(1+\zeta_1)^{\frac{f_{1-}}{\epsilon}-1}(1-\zeta_1)^{\frac{f_{2-}}{\epsilon}}|\zeta_1|^{\frac{f_{3-}}{\epsilon}} \\
\times \exp\biggl[\frac{1}{\epsilon^2}\int_{-1}^{\zeta_1}\int_{-1}^{\zeta_2}\biggl(&\frac{1}{(1+\zeta_2)(1-\zeta_3)} + \frac{1}{(1-\zeta_2)(1+\zeta_3)} + \frac{1}{\zeta_2\zeta_3}\biggr)Q(f(\zeta_3))\,d\zeta_3 d\zeta_2\biggr].
\end{aligned}$$

Invoking (3.13),

$$Q(f(\zeta)) \ge P_-(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2-}}{\epsilon}}|\xi|^{\frac{f_{3-}}{\epsilon}}$$

for $P_- < 0$ in the above, we obtain

$$\begin{aligned}
f_2(\xi) - f_{2-} \ge{}& \frac{-P_-}{\epsilon}\int_{-1}^{\xi}(1+\zeta_1)^{\frac{f_{1-}}{\epsilon}-1}(1-\zeta_1)^{\frac{f_{2-}}{\epsilon}}|\zeta|^{\frac{f_{3-}}{\epsilon}} \\
&\times \exp\biggl[\frac{1}{\epsilon^2}\int_{-1}^{\zeta_1}\int_{-1}^{\zeta_2}\biggl(\frac{1}{(1+\zeta_2)(1-\zeta_3)} + \frac{1}{(1-\zeta_2)(1+\zeta_3)} + \frac{1}{\zeta_2\zeta_3}\biggr) \\
&\times P_-(1+\zeta_3)^{\frac{f_{1-}}{\epsilon}}(1-\zeta_3)^{\frac{f_{2-}}{\epsilon}}|\zeta_3|^{\frac{f_{3-}}{\epsilon}}\,d\zeta_2 d\zeta_3\biggr]d\zeta_1 \\
={}& \frac{-P_-}{\epsilon}O(1)\int_{-1}^{\xi}(1+\zeta_1)^{\frac{f_{1-}}{\epsilon}-1}\exp\biggl\{\frac{P_-}{\epsilon^2}\int_{-1}^{\zeta_1}\int_{-1}^{\zeta_2}\biggl[O(1)(1+\zeta_3)^{\frac{f_{1-}}{\epsilon}-1} \\
&+ \frac{O(1)(1+\zeta_3)^{\frac{f_{1-}}{\epsilon}}}{1+\zeta_2} + O(1)(1+\zeta_3)^{\frac{f_{1-}}{\epsilon}}\biggr]d\zeta_2 d\zeta_3\biggr\}d\zeta_1
\end{aligned}$$

for $\xi \in [-1,-1/2]$. A further calculation of the above yields

$$f_2(\xi) - f_{2-} \ge -P_- A(1+\xi)^{\frac{f_{1-}}{\epsilon}}\exp\biggl[P_- B(1+\xi)^{\frac{f_{1-}}{\epsilon}+1}\biggr] \tag{3.45}$$

for $\xi \in [-1,-1/2]$, where $A, B > 0$ are constants depending only on f_- and ϵ. For any $N > 0$, let $P_- < -N^{\frac{\epsilon+f_{1-}}{\epsilon}}$. Then $(N/-P_-)^{\frac{\epsilon}{f_{1-}}} < 1/N$. Choosing

$$\xi_0 = -1 + \left(\frac{N}{-P_-}\right)^{\frac{\epsilon}{f_{1-}}} < -1 + \frac{1}{N},$$

we have

$$-P_-(1+\xi_0)^{\frac{f_{1-}}{\epsilon}} = N,$$
$$P_-(1+\xi_0)^{\frac{f_{1-}}{\epsilon}+1} = -N(1+\xi_0) > -N\cdot\frac{1}{N} = -1.$$

Then (3.45) reads

$$f_2(\xi_0) - f_{2-} \geq NA\exp(-B).$$

Since $P_- < 0$, the function $f_2(\xi; P_-)$ is increasing on $[-1, 0]$, and hence

$$f_2(0; P_-) - f_{2-} \geq NA\exp(-B)$$

if $P_- < -M^{\frac{\epsilon+f_{1-}}{\epsilon}}$. Thus

$$\lim_{P_-\to-\infty} f_2(0; P_-) = \infty. \qquad \square$$

Now we consider trajectories of (3.1) on $[0,1]$ with initial condition (3.3). Under the transformation

$$(\xi, f_1, f_2, f_3) \mapsto (-\xi, f_2, f_1, f_3), \tag{3.46}$$

problem (3.1), (3.3) becomes the initial-value problem of equation (3.1) with initial value $f(-1) = (f_{2+}, f_{1+}, f_{3+})$, which we have already studied. Our previous results then yield the following theorem.

THEOREM 3.9. *There is a $P_{0+} > 0$ such that the initial-value problem* (3.1), (3.3) *has a unique, positive solution $f^+(\xi; P_+)$ on* $[0,1]$ *for any given*

$$P_+ = \left.\frac{Q(f(\xi))}{(1-\xi)^{\frac{f_{2-}}{\epsilon}}}\right|_{\xi=1}, \quad P_+ \in (-\infty, P_{0+}].$$

Furthermore, we have the following:

(i)

$$f_1^+(0; P_{0+}) = 0. \tag{3.47}$$

(ii) $f^+(0; P_+)$ *is continuous with respect to* $P_+ \in (-\infty, P_{0+}]$.

(iii)

(3.48a)
$$C_+(f_{1+}, f_{2+}) := \left\{(f_1^+(0; P_+), f_2^+(0, ; P_+)) \in \mathbb{R}^2 : P_+ \in (-\infty, P_0)\right\}$$
$$\subset \left\{(f_1, f_2) \in \mathbb{R}^2 : f_1 \geq f_1^+(0; P_{0+}) = 0, f_2^+(0; P_{0+}) \geq f_2 > 0\right\}.$$

(iv)

$$\lim_{P_-\to-\infty} f_1^+(0; P_-) = +\infty. \tag{3.48b}$$

THEOREM 3.10. *Problem* (1.6) *always has a continuous positive solution.*

Proof. For any given $f_\pm > 0$, the curve $C_-(f_{1-}, f_{2-})$, parametrized as $(f_1(0; P_-),$ $f_2(0; P_-))$, $P_- \in (-\infty, P_0]$, runs from $f_2(0, P_0) = 0$ to $f_2(0; -\infty) = +\infty$ and $0 <$

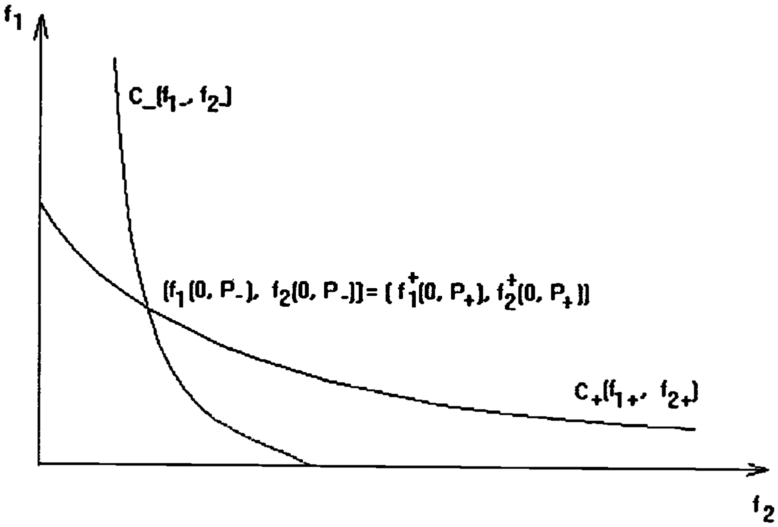


FIG. 1.

$f_1(0; P_-) \leq f_1(0, P_0)$ while $C_+(f_{1+}, f_{2+})$, parametrized by $(f_1^+(0; P_+),\ f_2^+(0; P_+))$, $P_+ \in (-\infty, P_{0+})$, runs from $f_1^+(0; P_{0+}) = 0$ to $f_1^+(0; -\infty) = +\infty$ with $0 < f_2^+(0; P_-) \leq f_2^+(0; P_{0+})$. By using the Lipschitz continuity of $(f_{1\pm}(0; P), f_{2\pm}(0; P))$ obtained in Corollary 3.7, we can prove that

$$C_-(f_{1-}, f_{2-}) \cap C_+(f_{1+}, f_{2+}) \neq \emptyset, \tag{3.49}$$

as shown in Fig. 1.

The proof of (3.49) is given in Lemma 3.11 below. Let

$$\begin{aligned} \big(f_1(0; P_-), f_2(0; P_-)\big) &= \big(f_1^+(0; P_+), f_2^+(0; P_+)\big) \\ &\in C_-(f_{1-}, f_{2-}) \cap C_+(f_{1+}, f_{2+}) \end{aligned} \tag{3.50}$$

for some $P_- \in (-\infty, P_0]$, $P_+ \in (-\infty, P_{0+}]$. Consider the function

$$f(\xi) = \begin{cases} f(\xi; f_{1-}, f_{2-}, P_-) & \text{if } \xi \in [-1, 0], \\ \\ f^+(\xi; f_{1+}, f_{2+}, P_+) & \text{if } \xi \in [0, 1], \end{cases} \tag{3.51}$$

which is continuous and positive on $[-1, 0) \cup (0, 1]$. Furthermore, the matching condition (3.50) says that f_1 and f_2 are continuous on $[-1, 1]$. Thus f_1 and f_2 are bounded and positive on $[-1, 1]$, and hence so is f_3. Since f is positive on $[-1, 0) \cup (0, 1]$, Lemma 3.1 (iii) and (iv) imply that the function $Q(f(\xi)$ is continuous at $\xi = 0$ and $Q|_{\xi=0} = 0$. Then $f_3(\xi)$ must be continuous at $\xi = 0$ as well. Then f defined by (3.51) satisfies the integral form of equation (3.1) and hence (3.1) itself. Thus f is a positive, continuous solution of (1.6). □

LEMMA 3.11.

$$C_-(f_{1-}, f_{2-}) \cap C_+(f_{1+}, f_{2+}) \neq \emptyset. \tag{3.52}$$

Proof. We recall from (3.36), (3.40), and (3.48a) that

$$\begin{aligned} C_-(f_{1-}, f_{2-}) &:= \big\{(f_1(0; f_{1-}, f_{2-}, P_-), f_2(0; f_{1-}, f_{2-}, P_-)) : P_- \in (-\infty, P_0)\big\} \\ &\subset \big\{(f_1, f_2) \in \mathbb{R}^2 : 0 < f_1 \le f_1(0; P_0), 0 = f_2(0; P_0) \le f_2\big\} \end{aligned} \tag{3.53}$$

and

$$\begin{aligned} C_+(f_{1+}, f_{2+}) &:= \big\{(f_1^+(0; P_+), f_2^+(0, ; P_+)) \in \mathbb{R}^2 : P_+ \in (-\infty, P_0)\big\} \\ &\subset \big\{(f_1, f_2) \in \mathbb{R}^2 : f_1 \ge f_1^+(0; P_{0+}) = 0, f_2^+(0; P_{0+}) \ge f_2 > 0\big\}. \end{aligned} \tag{3.54}$$

By Corollary 3.7, the curves $C_-(f_{1-}, f_{2-})$ and $C_+(f_{1+}, f_{2+})$ are Lipschitz continuous in $P_\pm$, respectively. For the curve $C_-(f_{1-}, f_{2-})$, we define the following subsets of $(-\infty, P_0]$:

$$\mathcal{A} := \cup\big\{(p, q) \subset (-\infty, P_0] : (f_!(0; p), f_2(0; p)) = (f_!(0; q), f_2(0; q))\big\}, \tag{3.55}$$

$$\mathcal{B} := (-\infty, P_0]\backslash\mathcal{A}. \tag{3.56}$$

We further define

$$t(P) := P_0 + \int_{P_0}^{P} \chi_{\mathcal{B}}(p)dp, \tag{3.57}$$

$$P(t) := \inf\{P : t(P) = t\}, \tag{3.58}$$

$$\big(\bar{f}_1(t), \bar{f}_2(t)\big) := \big(f_1(0; P(t)), f_2(0; P(t))\big). \tag{3.59}$$

From the definitions in (3.55)–(3.59), we can see that the curve $\big(\bar{f}_1(t), \bar{f}_2(t)\big)$ cannot be self-intersecting. Indeed, otherwise, there would be t_1 and t_2, $t_1 < t_2$, such that

$$\big(\bar{f}_1(t_1), \bar{f}_2(t_1)\big) = \big(\bar{f}_1(t_2), \bar{f}_2(t_2)\big) \tag{3.60}$$

and hence

$$(P(t_1), P(t_2)) \subset \mathcal{A}. \tag{3.61}$$

Then we have from (3.57) and (3.61) that

$$t_2 - t_1 = \int_{P(t_1)}^{P(t_2)} \chi_{\mathcal{B}}(p)dp = 0,$$

which is a contradiction. We notice from (3.57) that $t(P)$ is an increasing function of P and hence $P(t)$ is also an increasing function. We further claim that $\big(\bar{f}_1(t), \bar{f}_2(t)\big)$ is continuous in t. To prove the claim, we let t_0 be any point in the domain of $\big(\bar{f}_1(t), \bar{f}_2(t)\big)$. Since $P(t)$ is increasing, $P(t) \to p_++$ (or $p_-−$) as $t \to t_0+$ (or $t_0−$). This implies that

$$\lim_{t \to t_0+} \bar{f}_1(t) = \lim_{t \to t_0+} f_1(0; P(t)) = f_1(0; p_+) \tag{3.62}$$

and

$$\lim_{t \to t_0-} \bar{f}_1(t) = \lim_{t \to t_0-} f_1(0; P(t)) = f_1(0; p_-). \tag{3.63}$$

From the definition in (3.57), we have

$$t_0 = P_0 + \int_{P_0}^{p_-} \chi_{\mathcal{B}}(y)dy$$

and

$$t_0 = P_0 + \int_{P_0}^{p_+} \chi_{\mathcal{B}}(y)dy,$$

which yield

$$\int_{p_-}^{p_+} \chi_{\mathcal{B}}(y)dy = 0. \tag{3.64}$$

Consider the open set $\mathcal{A} \cap (p_-, p_+)$. It is a union of countably many disjoint open intervals

$$\mathcal{A} \cap (p_-, p_+) = \cup_{n=1}^{\infty} (p_n, q_n).$$

Then (3.64) is equivalent to

$$\sum_{n=1}^{\infty} (q_n - p_n) = p_+ - p_-. \tag{3.65}$$

From the definition of $\mathcal{A}$, it is clear that $f_1(0; p_n) = f_1(0; q_n)$. We consider $\cup_{n=1}^{N} (p_n, q_n)$ for large integer N. We assume without loss of generality—rearranging the notation if necessary—that $p_1 < q_1 < p_2 < q_2 < \cdots < q_N$. The difference of $f_1(0; p_-)$ and $f_1(0; p_+)$ satisfies

$$\begin{aligned}
&|f_1(0; p_-) - f_1(0; p_+)| \\
&\quad = \left| f_1(0; p_-) - \sum_{n=1}^{N} \big(f_1(0; p_n) - f_1(q_n)\big) - f_1(0; p_+) \right| \\
&\quad = \left| f_1(0; p_-) - f_1(0; p_1) + \sum_{n=1}^{N-1} \big(f_1(0; q_n) - f_1(0; p_{n+1})\big) + f_1(0; q_N) - f_1(0; p_+) \right| \\
&\quad \le O(1) \left(|p_1 - p_-| + \sum_{n=1}^{N-1} |q_n - p_{n+1}| + |p_+ - q_N| \right) \\
&\quad = O(1) \left(p_+ - p_- - \sum_{n=1}^{N} (q_n - p_n) \right) \\
&\quad \to 0 \quad \text{as } N \to \infty,
\end{aligned} \tag{3.66}$$

where we used the Lipschitz continuity of f_1 on P and (3.65). This proves the continuity of $\bar{f}_1(t)$. We can prove the continuity of $\bar{f}_2(t)$ similarly.

Now we claim that $\bar{f}_2(t) \to \infty$ as $t \to \inf(\text{domain of definition of } (\bar{f}_1, \bar{f}_2))$. To this end, it suffices to prove that there is a sequence $\{P_n\} \subset \mathcal{B}$ such that $P_n \to -\infty$ as $n \to \infty$ since $f_2(0; P) \to \infty$ as $P \to -\infty$. Indeed, otherwise, by the definition of $\mathcal{B}$, there would be a sequence $\{\bar{P}_n\} \subset \mathcal{A}$ such that $\bar{P}_n \to -\infty$ as $n \to \infty$ and $f_2(0; \bar{P}_n)$

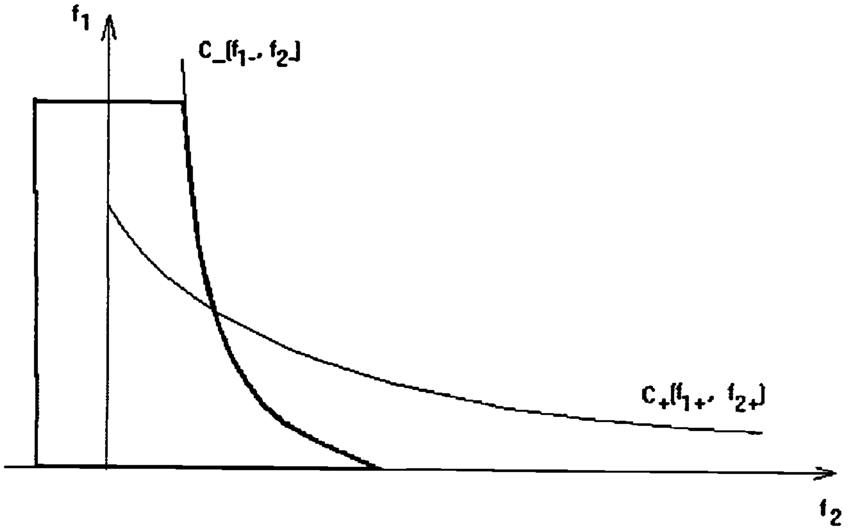


Fig. 2.

is bounded, which is a contradiction to the fact that $f_2(0; P) \to \infty$ as $P \to \infty$. The claim is proved.

The domain of definition of $(\bar{f}_1, \bar{f}_2)$ is the range of $t(P)$ defined in (3.57), which is an interval due to the continuity of $t(P)$. As P runs from P_0 to $-\infty$, the function $t(P)$ decreases from P_0 to inf(domain of definition of $(\bar{f}_1, \bar{f}_2)$) and $\bar{f}_2(t)$ runs from 0 to ∞. There is at least one point $T \in$(domain of definition of $(\bar{f}_1, \bar{f}_2)$) such that $\bar{f}_2(T) = 2f_{2+}(0; P_{0+})$. Let $\mathcal{T}$ be the set of all of these T's and $T_1 \in \mathcal{T}$ be such that $\min_{T \in \mathcal{T}} \bar{f}_1(T)$ is attained. We define a closed curve

$$\begin{aligned}\mathcal{J} := &\{(\bar{f}_1, \bar{f}_2)(t) : t \in [T_1, P_0]\}\\ &\cup \{(f_1, f_2) : f_2 = \bar{f}_2(T_1),\ -1 \le f_1 \le \bar{f}_1(T_1)\}\\ &\cup \{(-1, f_2) : 0 \le f_2 \le \bar{f}_2(T_1)\} \cup \{(f_1, 0) : -1 \le f_1 \le f_1(0; P_0)\},\end{aligned} \tag{3.67}$$

which is depicted in Fig. 2.

From the definition in (3.67), the definition of T_1, and (3.53), we can see that the curve $\mathcal{J}$ is a simple closed curve. Jordan's curve theorem then states that the curve $\mathcal{J}$ divides the whole (f_1, f_2)-plane into two components, an interior component and an exterior component. The curve $C_+(f_{1+}, f_{2+})$ has a point $(0, f_2^+(0; P_{0+}))$ inside the interior component and a point in the exterior component since $(f_1^+(0; P) \to \infty$ as $P \to \infty$. Thus the curve $C_+(1+, f_{2+})$ must intersect the boundary of the interior component, which is the curve $\mathcal{J}$. From (3.54), we see that $C_+(f_{1+}, f_{2+})$ can intersect $\mathcal{J}$ only at the $\{(\bar{f}_1, \bar{f}_2)(t) : t \in [T_1, P_0]\}$ part of $\mathcal{J}$, which is a portion of the curve $C_-(f_{1-}, f_{2-})$. Thus we obtain the desired conclusion. □

The following corollary justifies the passing of the fluid dynamic limit in (1.5) to obtain the limit equation (1.2).

Corollary 3.12. *For any Maxwellian Riemann data* (1.4)*, there is a solution of the equation* (1.4)–(1.5)*, $f^\epsilon(x/t)$. Further, there is a sequence of solutions of* (1.4)–(1.5)*, $\{f^{\epsilon_n}\}$, $\epsilon_n \to 0+$, such that $f^{\epsilon_n} \to f(x/t)$ almost everywhere $(x, t) \in \mathbb{R} \times \mathbb{R}_+$, where $f(x/t)$ is a solution of the limit equation* (1.2), (1.4).

Proof. The first assertion is a restatement of Theorem 3.10. In [ST], Slemrod and

Tzavaras proved that the total variation of solutions of (1.4)–(1.5) bounded uniformly in ϵ. Thus there is a sequence of solutions of (1.4)–(1.5), $\{f^{\epsilon_n}\}$, $\epsilon_n \to 0+$, such that $f^{\epsilon_n} \to f(x/t)$ almost everywhere $(x,t) \in \mathbb{R} \times \mathbb{R}_+$. It is clear that the limit function is a weak solution of (1.2), (1.4). ▯

With help from Lemma 3.1, we can improve the result of Lemma 2.2 as follows.

Theorem 3.13. *Let f be a positive solution of* (3.1) *with boundary condition* (3.2)–(3.3). *Then*

$$f_1(\xi) - f_{1+} = O(1)(1-\xi)^{\frac{f_{2+}}{\epsilon}} \quad \textit{for } \xi \textit{ near } \xi = 1, \tag{3.68}$$

$$f_2(\xi) - f_{2-} = O(1)(\xi+1)^{\frac{f_{1-}}{\epsilon}} \quad \textit{for } \xi \textit{ near } \xi = -1, \tag{3.69}$$

and

$$f_3(\xi) - f_3(0) = O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}} \quad \textit{for } \xi \textit{ near } \xi = 0, \tag{3.70}$$

$$Q(f\xi)) = O(1)(\xi+1)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}}. \tag{3.71}$$

Proof. From (3.5a, b), we know that

$$Q(f(\xi)) = O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}}. \tag{3.72}$$

Then by integrating $(3.1)_3$, we obtain

$$f_3(\xi) - f_3(0) = \int_0^\xi \frac{Q(f(\zeta)}{\epsilon\zeta} d\zeta = O(1)|\xi|^{\frac{\min(f_{3-}, f_3(0), f_{3+})}{\epsilon}},$$

which is (3.70).

To prove (3.68), we start with (3.4a),

$$\begin{aligned} Q &= C(1+\xi)^{\frac{f_{1-}}{\epsilon}}(1-\xi)^{\frac{f_{2+}}{\epsilon}}|\xi|^{\frac{\min(f_{3-}, f_3(0-))}{\epsilon}} \\ &\quad \times \exp\left[\frac{1}{\epsilon}\int_{-1}^{\xi}\left(\frac{f_1(\zeta) - f_{1-}}{\zeta+1} + \frac{f_2(\zeta) - f_{2+}}{\zeta - 1} + \frac{f_3(\zeta) - \min(f_{3-}, f_3(0-)}{\zeta}\right)d\zeta\right], \end{aligned} \tag{3.73}$$

which holds for $\xi \in [-1,0)$. If $C \geq 0$ and hence $Q \geq 0$ on $[-1,0)$, then f_2 and f_3 are decreasing on $[-1,0)$, and hence

$$f_1' = \frac{Q}{\epsilon(1-\xi)} \leq \frac{f_3^2(\xi)}{\epsilon(1-\xi)} \leq \frac{f_{3-}^2}{\epsilon}.$$

Thus all of the terms in the integrand of (3.71) are either negative or finite, which implies that

$$Q(f(\xi)) = O(1)(1+\xi)^{\frac{f_{1-}}{\epsilon}} \quad \text{for } \xi \in [-1,). \tag{3.74}$$

This together with $(3.1)_2$ yields that

$$f_2(\xi) - f_{2-} = \int_{-1}^{\xi} \frac{-Q(f(\zeta))}{\epsilon(\zeta+1)} = O(1)(1+\xi)^{\frac{f_{1-}}{\epsilon}},$$

which proves (3.68). Similarly, we can prove that

$$Q(f(\xi)) = O(1)(1-\xi)^{\frac{f_{2+}}{\epsilon}} \quad \text{for } \xi \in (0,1] \tag{3.75}$$

and thus (3.69). Combining (3.72), (3.74), and (3.75), we obtain (3.71). ▯

Acknowledgments. I would like to thank Professor Marshall Slemrod for his suggestions. I am grateful to Professor Tai-Ping Liu for his hospitality and for many valuable discussions while I was visiting Stanford University, where this work was done. I also wish to thank the referee, whose comments helped me to revise this paper. My thanks also go to Professor Andrew Vogt of Georgetown University for our discussion of the proof of Lemma 3.11.

REFERENCES

[B] J. E. Broadwell, *Shock structure in a simple discrete velocity gas*, Phys. Fluids, 7 (1964), pp. 1243–1247.

[BGL] C. Bardos, F. Golse, and C. D. Levermore, *Fluid dynamic limits of kinetic equations* II: *Convergence proofs for the Boltzmann equation*, Comm. Pure Appl. Math., 46 (1993), pp. 667–753.

[Ca] R. E. Caflisch, *The fluid dynamic limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math., 33 (1980), pp. 651–666.

[Ce] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer-Verlag, New York, 1988.

[CL] R. E. Caflisch and T.-P. Liu, *The stability of shock waves for the Broadwell equations*, Comm. Math. Phys., 114 (1988), pp. 103–130.

[ChL] Q.-C. Chen and T.-P. Liu, *Zero relaxation and dissipation limits for hyperbolic conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 755–781.

[CLL] Q.-C Chen, D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[CP] R. E. Caflisch and G. C. Papanicolaou, *The fluid dynamic limit of a nonlinear model Boltzmann equation*, Comm. Pure Appl. Math., 32 (1979), pp. 589–616.

[KM] S. Kawashima and A. Matsumura, *Asymptotic stability of traveling wave solutions of a system of one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.

[LX] J.-G. Liu and Z. Xin, *Boundary layer behavior in the fluid dynamic limit for a nonlinear model of Boltzmann equation*, Arch. Rat. Mech. Anal., 135 (1996), pp. 61–105.

[Ma] A. Matsumura, *Asymptotic toward rarefaction wave of solutions of the Broadwell model of a discrete velocity gas*, Japan J. Appl. Math. 4 (1987), pp. 489–502.

[PI] T. Platkowski and R. Illner, *Discrete velocity models of the Boltzmann equation: A survey of the mathematical aspects of the theory*, SIAM Rev., 30 (1988), pp. 213–255.

[ST] M. Slemrod and A. Tzavaras, *Self-similar fluid dynamic limit for the Broadwell system*, Arch. Rat. Mech. Anal., 122 (1993), pp. 353–392.

[T] A. Tzavaras, *Wave structure induced by fluid dynamic limits in the Broadwell model*, Arch. Rat. Mech. Anal., 127 (1994), pp. 361–387.

[X] Zhouping Xin, *The fluid dynamic limit for the Broadwell model of the nonlinear Boltzmann equation in the presence of shocks*, Comm. Pure Appl. Math., 44 (1991), pp. 679–741.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 852–866, July 1997

RANGE OF THE RADON TRANSFORM ON FUNCTIONS WHICH DO NOT DECAY FAST AT INFINITY*

ALEXANDER I. KATSEVICH[†]

Abstract. Let an integer $m \geq 0$ be fixed. Let $\mathcal{X}_m$ be the space of functions $f \in C^\infty(\mathbb{R}^n)$ that admit an asymptotic expansion $f(r\beta) \sim \sum_{k=m}^{\infty} \psi_k(\beta)/r^{n+k}, r \to \infty$, $\psi_k \in C^\infty(S^{n-1})$, and the expansion can be differentiated with respect to $x = r\beta$ any number of times. In this paper, we derive a precise characterization of the range of the Radon transform R acting on $\mathcal{X}_m$; that is, we explicitly describe the space $\mathcal{Z}_m = R\mathcal{X}_m$. The conditions which describe the space $\mathcal{Z}_m$ are easily verifiable.

Key words. range, Radon transform, moment conditions, asymptotic expansion

AMS subject classifications. 44A12, 65R10, 92C55

PII. S0036141095289518

1. Introduction and statement of main result.

Consider the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ consisting of $C^\infty(\mathbb{R}^n)$ functions that decay with all derivatives faster than any power of $|x|^{-1}$ as $|x| \to \infty$. If $f \in \mathcal{S}(\mathbb{R}^n)$, then Rf, the Radon transform of f, is defined by the formula

$$Rf(\alpha, p) = \int_{\mathbb{R}^n} f(x)\delta(\alpha \cdot x - p)dx, \quad \alpha \in S^{n-1}, \quad p \in \mathbb{R}. \tag{1.1}$$

Here S^{n-1} is the unit sphere in $\mathbb{R}^n$ and δ is the one-dimensional delta function. If f is a distribution, then Rf is defined via duality $(Rf, \varphi) = (f, R^*\varphi)$, where R^* is the operator dual to R,

$$R^*\varphi(x) = \int_{S^{n-1}} \varphi(\alpha, \alpha \cdot x)d\alpha,$$

and the test functions φ run through a suitable space (see, e.g., [8], [4], and [12]).

Let $\mathcal{S}_{em}(Z)$, $Z := S^{n-1} \times \mathbb{R}$, be the space of $C^\infty(Z)$ functions $g(\alpha, p)$ that satisfy the following conditions:

1. g is even: $g(\alpha, p) = g(-\alpha, -p)$, $(\alpha, p) \in Z$;
2. $g(\alpha, p)$ and all of its derivatives with respect to α and p decay faster than any power of p^{-1} as $p \to \infty$; and
3.

$$\int_{-\infty}^{\infty} g(\alpha, p)p^j dp = \mathcal{P}_j(\alpha), \quad j = 0, 1, 2, \ldots, \tag{1.2}$$

where $\mathcal{P}_j(x), x \in \mathbb{R}^n$, is a homogeneous polynomial of degree j and $\mathcal{P}_j(\alpha)$ is the restriction of $\mathcal{P}_j$ to S^{n-1}.

*Received by the editors July 26, 1995; accepted for publication (in revised form) March 12, 1996. This research was performed under the auspices of the U.S. Department of Energy. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sima/28-4/28951.html

[†]Los Alamos National Laboratory, CIC-3 Division, Los Alamos, NM 87545. Current address: Mathematics Department, University of Central Florida, Orlando, FL 32816-1364 (akatsevi@pegasus.cc.ucf.edu).

The conditions in (1.2) are known as the "moment conditions." It was proved in [3] (see also [2] and [8]) that $R : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}_{em}(Z)$ is an isomorphism. This result gives a complete description of the range of R on the space of smooth rapidly decaying functions. Subsequently, ranges of R on spaces of compactly supported distributions $\mathcal{E}'$ and rapidly decaying distributions $\mathcal{O}'_c$ have been described [5], [6], [7]. Basically, these results state that the range of R on $\mathcal{E}'(\mathbb{R}^n)(\mathcal{O}'_c(\mathbb{R}^n))$ consists of $g \in \mathcal{E}'(Z)(\mathcal{O}'_c(Z))$, which are even and satisfy the moment conditions in (1.2) in a generalized sense. In case of $\mathcal{O}'_c(\mathbb{R}^n)$, it is necessary to assume in addition that g is C^∞ in the α variable [6]. Range theorems on the space of compactly supported H^s-functions have been given in [9], [13], [10], and [11]. In this case, functions from the range of R belong to the corresponding Sobolev space $H^{s+(n-1)/2}(Z)$, are even and compactly supported, and satisfy the moment conditions in (1.2).

A preimage under the Radon transform of $\mathcal{S}(Z)$ even functions that satisfy none or finitely many of the conditions in (1.2) was studied in [14]. Suppose that $g \in \mathcal{S}(Z)$ and g is even. Among other results, it was proved in [14] that the more moment conditions that g satisfies, the faster $f := R^{-1}g$ decays at infinity. Moreover, if g does not satisfy all of the moment conditions, then the decay rate of f is only polynomial. A related result was obtained in [15], where a partial range theorem for R on functions which decay only polynomially was proved. In particular, it was proved in [15] that if $g \in C^r(Z), r < \infty$, satisfies finitely many moment conditions and derivatives of $g(\alpha, p)$ with respect to α decay sufficiently fast as $p \to \infty$ (but not necessarily exponentially fast), then there exists a function f such that $Rf = g$ and the inversion formula $f = (1/2(2\pi)^{n-1})\Lambda^{n-1}R^*g$ holds. Here Λ is the Calderon operator defined by the equation $\Lambda^2 = -\Delta$, where Δ is the Laplacian.

In this paper, we give a precise characterization of the range of R on a certain class of C^∞ functions that decay only polynomially at infinity. Our main result is the following.

THEOREM. *Fix any integer $m \geq 0$. Let $\mathcal{X}_m$ be the space of functions $f \in C^\infty(\mathbb{R}^n)$ that admit an asymptotic expansion*

$$f(r\beta) \sim \sum_{k=m}^{\infty} \frac{\psi_k(\beta)}{r^{n+k}}, \quad r \to \infty, \quad \psi_k \in C^\infty(S^{n-1}), \tag{1.3}$$

and the expansion can be differentiated with respect to $x = r\beta$ any number of times. Let $\mathcal{Z}_m$ be the space of even functions $g \in C^\infty(Z)$ such that

(a) *g satisfies the first m moment conditions* (1.2), *$j = 0, 1, \ldots, m-1$, and*

(b) *g admits an asymptotic expansion*

$$g(\alpha, p) \sim \sum_{k=m}^{\infty} \frac{Q_k(\alpha) + b_k(\alpha)}{p^{k+1}}, \quad p \to +\infty, \tag{1.4}$$

where

$$b_k \in C^\infty(S^{n-1}), \quad b_k(-\alpha) = (-1)^{k+1} \quad b_k(\alpha), \tag{1.5a}$$

$$Q_k \text{ is a homogeneous polynomial of degree } k, \tag{1.5b}$$

and the expansion can be differentiated with respect to p any number of times.

Then $R\mathcal{X}_m = \mathcal{Z}_m$.

Remark. Expansion (1.3) means that $f(r\beta)-\sum_{k=m}^{K-1}\psi_k(\beta)r^{-(n+k)}=O(r^{-(n+K)})$ as $r\to+\infty$ uniformly in $\beta\in S^{n-1}$ for each $K\geq m+1$. Expansion (1.4) means that $g(\alpha,p)-\sum_{k=m}^{K-1}(Q_k(\alpha)+b_k(\alpha))/p^{k+1}=O(p^{-(K+1)})$ as $p\to+\infty$ uniformly in $\alpha\in S^{n-1}$ for each $K\geq m+1$.

We see that the theorem presented is a generalization of the classical range theorem because $\mathcal{S}(\mathbb{R}^n)=\cap_{m\geq0}\mathcal{X}_m$ and $\mathcal{S}_{em}(Z)=\cap_{m\geq0}\mathcal{Z}_m$. One can also see that the conditions on the range of the Radon transform are easily verifiable.

For convenience, we introduce another space of functions. Let $\mathcal{Y}_m$ be the space of even functions $g\in C^\infty(Z)$ such that g admits an asymptotic expansion

$$g(\alpha,p)\sim\sum_{k=m}^{\infty}g_k(\alpha,p),\quad p\to\infty, \tag{1.6}$$

and the expansion can be differentiated with respect to p any number of times. In (1.6), the functions $g_k\in C^\infty(Z)$, $k\geq m$, are even and satisfy the following assumptions:

(1) g_k satisfies the first k moment conditions in (1.2), $j=0,1,\dots,k-1$;

(2) g_k can be represented as

$$\begin{gathered} g_k(\alpha,p)=\frac{Q_k(\alpha)+b_k(\alpha)}{p^{k+1}},\quad p>1,\\ b_k\in C^\infty(S^{n-1}),\quad b_k(-\alpha)=(-1)^{k+1}b_k(\alpha), \end{gathered} \tag{1.7}$$

where Q_k is a homogeneous polynomial of degree k; and

(3) for all $K\geq m$, the difference $g-\sum_{k=m}^{K-1}g_k$ satisfies the first K moment conditions in (1.2), $j=0,1,\dots,K-1$.

Proposition. $\mathcal{Z}_m=\mathcal{Y}_m$.

Proof. The inclusion $\mathcal{Y}_m\subset\mathcal{Z}_m$ is obvious. To prove the inclusion $\mathcal{Z}_m\subset\mathcal{Y}_m$, we have to fix any $g\in\mathcal{Z}_m$ and find functions g_k with properties (1)–(3). Let $h_k\in C^\infty(Z)$, $k\geq m$, be any even functions such that

$$h_k(\alpha,p)=\frac{Q_k(\alpha)+b_k(\alpha)}{p^{k+1}},\quad p>1, \tag{1.8a}$$

(1.8b) h_k satisfies the first k moment conditions.

The functions h_k can be constructed, for example, as follows:

$$h_k(\alpha,p)=Q_k(\alpha)A_k(p)+b_k(\alpha)B_k(p),$$

where $A_k,B_k\in C^\infty(\mathbb{R})$ are chosen so that

$$A_k(p)=B_k(p)=\frac{1}{p^{k+1}},\quad p>1;$$

$$A_k(-p)=(-1)^kA_k(p),\qquad B_k(-p)=(-1)^{k+1}B_k(p),\quad p\in\mathbb{R};$$

$$\int_{-\infty}^{\infty}A_kp^j\,dp=\int_{-\infty}^{\infty}B_kp^j\,dp=0,\quad 0\leq j\leq k-1.$$

Fix any $P_k \in C_0^\infty([-1,1])$ with the properties

$$P_k(-p) = (-1)^k P_k(p), \qquad \int_{-\infty}^{\infty} P_k(p)p^j dp = \begin{cases} 0, & j = 0,1,\dots,k-1, \\ 1, & j = k, \end{cases} \tag{1.9}$$

and define

(1.10)

$$g_k(\alpha,p) := h_k(\alpha,p) + P_k(p) \cdot \int_{-\infty}^{\infty} \left(g(\alpha,t) - \sum_{j=m}^{k-1} g_j(\alpha,t) - h_k(\alpha,t) \right) t^k dt, \quad k \geq m.$$

If $k = m$, the summation on the right-hand side of (1.10) disappears.

Let us check that the functions g_k defined in (1.10) have all of the required properties. Equations (1.4) and (1.8a) ensure that (1.6) is satisfied. Since (1.4) can be differentiated with respect to p, expansion (1.6) can be differentiated with respect to p as well. Using (1.10), it is easy to check that g_k, $k \geq m$, are $C^\infty(Z)$ and even. Property (1) follows immediately from (1.8b) and (1.9). Property (2) follows from (1.8a). Property (3) can be checked by induction.

Let us denote

$$\Delta g_m := g, \qquad \Delta g_k := g - \sum_{k=m}^{k-1} g_j, \quad k \geq m+1.$$

We have to show that Δg_k, $k \geq m$, satisfies k moment conditions. If $k = m$, then $\Delta g_m = g$ and by assumption (a) of the theorem, Δg_m satisfies m moment conditions. Now let $k > m$. Suppose that Δg_k satisfies k moment conditions. We have to show that Δg_{k+1} satisfies $k+1$ moment conditions. From the relation $\Delta g_{k+1} = \Delta g_k - g_k$, it is obvious that Δg_{k+1} satisfies k moment conditions. Let us check the $(k+1)$st moment condition. Using (1.9) and (1.10), which can be rewritten as

$$g_k(\alpha,p) := h_k(\alpha,p) + P_k(p) \cdot \int_{-\infty}^{\infty} (\Delta g_k(\alpha,t) - h_k(\alpha,t))\, t^k dt, \quad k \geq m,$$

we get

$$\begin{aligned} \int_{-\infty}^{\infty} \Delta g_{k+1} p^k dp &= \int_{-\infty}^{\infty} (\Delta g_k - g_k) p^k dp \\ &= \int_{-\infty}^{\infty} (\Delta g_k - h_k) p^k dp - \int_{-\infty}^{\infty} P_k(p) p^k dp \int_{-\infty}^{\infty} (\Delta g_k - h_k) t^k dt = 0. \end{aligned}$$

Since the right-hand side of the last equation is a homogenous polynomial of any degree, the desired assertion is proved. □

In view of the proposition, the assertion $R\mathcal{X}_m = \mathcal{Y}_m$ is equivalent to the theorem. In section 2, a proof of the inclusion $R\mathcal{X}_m \subset \mathcal{Y}_m$ is given. In section 3, we prove the inclusion $\mathcal{Y}_m \subset R\mathcal{X}_m$. Two auxiliary lemmas are proved in section 4.

2. Proof of the inclusion $R\mathcal{X}_m \subset \mathcal{Y}_m$. Let $m \geq 0$ be fixed. Fix any $f \in \mathcal{X}_m$ and show that $Rf \in \mathcal{Y}_m$. Define g by

$$g(\alpha,p) := Rf(\alpha,p) = \int_0^\infty \int_{S_{\alpha^\perp}^{n-2}} f(p\alpha + t\omega) d\omega\, t^{n-2} dt, \tag{2.1}$$

where $S^{n-2}_{\alpha^\perp}$ is the unit sphere in the hyperplane passing through the origin perpendicular to α. Expansion (1.3) ensures that the above integral converges absolutely. Since derivatives of f decay at least as fast as f, one can differentiate with respect to p under the integral sign in (2.1) any number of times. Now let us check that g can be differentiated with respect to α. Fix any $\alpha \in S^{n-1}$, and let $\omega \perp \alpha$. It is sufficient to check that the function $g(\sqrt{1-|\omega|^2}\alpha+\omega, p)$ can be differentiated with respect to ω at $|\omega| = 0$. Let us represent $x \in \mathbb{R}^n$ as follows: $x = s\alpha + y$, $y \perp \alpha$. Differentiating the identity

$$g(\sqrt{1-|\omega|^2}\alpha+\omega, p) = \int_{\mathbb{R}^{n-1}} f\left(\frac{p-\omega\cdot y}{\sqrt{1-|\omega|^2}}\alpha + y\right) dy$$

with respect to $\omega_k, 1 \le k \le n-1$, and setting $|\omega| = 0$, we get

$$\begin{aligned}\frac{\partial}{\partial \omega_k} g(\sqrt{1-|\omega|^2}\alpha+\omega, p)\bigg|_{|\omega|=0} &= -\int_{\mathbb{R}^{n-1}} y_k \frac{\partial}{\partial s} f(s\alpha+y)\bigg|_{s=p} dy \\ &= -R\left[y_k \frac{\partial}{\partial s} f(s\alpha+y)\right](\alpha, p).\end{aligned}$$

Since $|f(x)| \le O(|x|^{-n})$, $|\frac{\partial}{\partial s} f(s\alpha+y)| \le O((s^2+|y|^2)^{-(n+1)/2})$, and therefore $|y_k \frac{\partial}{\partial s} f(s\alpha+y)| \le O((s^2+|y|^2)^{-n/2})$, the integrals on the right-hand sides of the last two equations converge absolutely, and taking the derivative under the integral sign is justified. As a generalization, we get

$$P_m(\partial_\omega) g(\sqrt{1-|\omega|^2}\alpha+\omega, p)\bigg|_{|\omega|=0} = (-1)^m R\left[P_m(y)\frac{\partial^m}{\partial s^m} f(s\alpha+y)\right](\alpha, p),$$

where $P_m(\omega) = P_m(\omega_1, \dots, \omega_{n-1})$ is a homogeneous polynomial of degree m. By assumption, $|\frac{\partial^m}{\partial s^m} f(s\alpha+y)| \le O((s^2+|y|^2)^{-(n+m)/2})$, and therefore $R\big(P_m(y)\frac{\partial^m}{\partial s^m} f(s\alpha+y)\big)$ is well defined for any $m \ge 0$. This shows that $g(\alpha, p)$ is infinitely differentiable with respect to α. Therefore, $g \in C^\infty(Z)$. Clearly, g is even.

Let f_k be any $C^\infty(\mathbb{R}^n)$ function such that $f_k(r\beta) = r^{-(n+k)}\psi_k(\beta)$, $r > 1$ (see (1.3)). Denote $g_k = Rf_k$. Analogously, we check that $g_k \in C^\infty(Z)$ and g_k is even. The functions $f_k(r\beta)r^j$, $j = 0, 1, \dots, k-1$, are absolutely integrable, and therefore the identity

$$\int_{\mathbb{R}^n} f(x)(\alpha\cdot x)^j dx = \int_{-\infty}^{\infty} g(\alpha, p)p^j dp, \quad j = 0, 1, \dots, k-1, \quad g = Rf, \tag{2.2}$$

which follows from the Fubini theorem, implies that g_k satisfies assumption (1) (see below (1.6)). Representing f_k and g_k as $f_k(r\beta) = \sum_{l=0}^\infty R_{kl}(r)Y_l(\beta)$, where Y_l is the spherical harmonic of degree l, $R_{kl}(r) = c_{kl}r^{-(n+k)}$ for $r > 1$, and $g_k(\alpha, p) = \sum_{l=0}^\infty P_{kl}(p)Y_l(\alpha)$, using (1.1) and the Funk–Hecke theorem [12, pp. 18 and 19], we find for $p > 1$ that

$$\begin{aligned}&R\big(R_{kl}(r)Y_l(\beta)\big) \\ &\quad= \gamma_{nl}c_{kl}Y_l(\alpha)\int_p^\infty r^{-(n+k)}C_l^{(\frac{n-2}{2})}(p/r)(1-(p/r)^2)^{\frac{n-3}{2}}r^{n-2}dr \\ &\quad= \gamma_{nl}c_{kl}\frac{Y_l(\alpha)}{p^{k+1}}\int_0^1 t^k C_l^{(\frac{n-2}{2})}(t)(1-t^2)^{\frac{n-3}{2}}dt \\ &\quad= 0 \quad \text{if } k < l \text{ and } k+l \text{ is even,}\end{aligned} \tag{2.3}$$

where γ_{nl} are some nonzero constants. Thus we have verified assumption (2), and we have showed that the kth term of the expansion for f is transformed into the kth term of series (1.6).

Let $f_{\Sigma K} := \sum_{k=m}^{K-1} f_k$ be the sum of the first $K - m$ terms of expansion (1.3). Define $g_{\Sigma K} := Rf_{\Sigma K}$. Since $f(r\beta) - f_{\Sigma K}(r\beta) = O(r^{-(n+K)})$, equation (2.2) implies that assumption (3) is also verified. Moreover, using (2.1), we get

$$\begin{aligned} |g(\alpha,p) - g_{\Sigma K}(\alpha,p)| &\le \int_0^\infty \int_{S^{n-2}_{\alpha^\perp}} |f(p\alpha + t\omega) - f_{\Sigma K}(p\alpha + t\omega)| d\omega\, t^{n-2} dt \\ &\le \int_0^\infty O\Big((p^2+t^2)^{-\frac{n+K}{2}}\Big) t^{n-2} dt = O\big(p^{-(K+1)}\big). \end{aligned}$$

Therefore, the formal series (1.6), which was obtained by taking the Radon transform of (1.3) term by term, is, in fact, the asymptotic expansion. Let us show that (1.6) can be differentiated with respect to p. Let $Q_s(x)$ be a homogeneous polynomial of degree $s \ge 0$. By assumption, the expansion $f(r\beta) \sim \sum_{k=m}^\infty f_k(r\beta)$ can be differentiated any number of times with respect to $x = r\beta$. Then $Q_s(\partial_x) f(r\beta) \sim \sum_{k=m}^\infty Q_s(\partial_x) f_k(r\beta)$. Taking the Radon transform on both sides and using the identity $R(Q_s(\partial_x)f) = Q_s(\alpha)\partial_p^s g(\alpha,p), g = Rf$ (see [14, Lemma 4.3] and [12, equation (2.1.15)]), we get by the already proved part of the theorem that

$$Q_s(\alpha)\partial_p^s g(\alpha,p) \sim \sum_{k=m}^\infty R(Q_s(\partial_x)f_k) = \sum_{k=m}^\infty Q_s(\alpha)\partial_p^s g_k(\alpha,p).$$

In view of the remark following the theorem, this means that

$$Q_s(\alpha)\bigg(\partial_p^s g(\alpha,p) - \sum_{k=m}^{K-1} \partial_p^s g_k(\alpha,p)\bigg) = O(\partial_p^s g_K) = O(p^{-(K+s+1)})$$

uniformly in α. The last equation holds for any homogeneous polynomial Q_s. Substituting two polynomials Q_s with disjoint zeros (e.g., $Q_s(\alpha) = \alpha_i^s$ and $Q_s(\alpha) = \alpha_j^s, i \ne j$), we see that

$$\partial_p^s g(\alpha,p) - \sum_{k=m}^{K-1} \partial_p^s g_k(\alpha,p) = O(p^{-(K+s+1)})$$

uniformly in α. Therefore, expansion (1.6) can be differentiated with respect to p. Thus we have checked that $Rf \in \mathcal{Y}_m$ if $f \in \mathcal{X}_m$. □

3. Proof of the inclusion $\boldsymbol{\mathcal{Y}_m \subset R\mathcal{X}_m}$. Let us show that if $g \in \mathcal{Y}_m$, then $f = R^{-1}g \in \mathcal{X}_m$. Let $\mathcal{F}$ denote the Fourier transform in $\mathbb{R}^n$ and F denote the one-dimensional Fourier transform acting on the p variable. The Fourier slice theorem [12, p. 15] asserts that $\mathcal{F}f = F(Rf)$ if $f \in \mathcal{S}(\mathbb{R}^n)$. In [14], it was shown that this relation holds for a much larger class of functions, which includes, in particular, functions of the type of (1.3) (see Lemma 4.5 in [14]). This yields two convenient formulas: $R = F^{-1}\mathcal{F}$ and $R^{-1} = \mathcal{F}^{-1}F$. Denote $\tilde g(\alpha,\lambda) = F_{p\to\lambda} g(\alpha,p)$. Using the formula for R^{-1} and writing $\mathcal{F}^{-1}$ in spherical coordinates, define

$$(3.1) \quad f(r\beta) := \mathcal{F}^{-1}Fg = \frac{1}{(2\pi)^n}\int_{S^{n-1}}\int_0^\infty \tilde g(\alpha,\lambda) e^{-ir\lambda(\alpha\cdot\beta)} \lambda^{n-1} d\lambda\, d\alpha, \quad g \in \mathcal{Y}_m.$$

Let us compute $R^{-1}g_k$, where g_k is a term from expansion (1.6). Without loss of generality, we may assume that g_k satisfies the first k moment conditions with the corresponding homogeneous polynomials being identically equal zero. Indeed, it is well known (see, e.g., [14, Lemma 7.4] and [12, Exercise 3.1.1 on p. 69]) that there exists $\varphi_k \in C_0^\infty(\mathbb{R}^n)$, $\varphi_k(x) = 0$ for $|x| \geq 1$, such that

$$\int_{-\infty}^{\infty} (g_k(\alpha,p) - R\varphi_k(\alpha,p))p^j\,dp = 0, \quad j = 0, 1, \dots, k-1.$$

Replacing g_k by $g_k - R\varphi_k$ in (1.6), we see that all of the assumptions are still satisfied, and expansion (1.6) remains valid. Let $P_{kl} \in C^\infty(\mathbb{R})$ be a function such that

$$P_{kl}(p) = p^{-(k+1)}, \quad p > 1; \qquad P_{kl}(p) = -p^{-(k+1)}, \quad p < -1; \tag{3.2}$$

$$\int_{-\infty}^{\infty} P_{kl}(p)p^j\,dp = 0, \quad j = 0, 1, \dots, k-1. \tag{3.3}$$

Using (1.7), we get

$$g_k(\alpha,p) = \sum_{l \leq k,\ l+k \text{ even}} a_{kl} P_{kl}(p) Y_l(\alpha) + \Delta g_k(\alpha,p), \tag{3.4}$$

where the constants a_{kl} are determined from the equation $\sum_{l \leq k,\ l+k \text{ even}} a_{kl} Y_l(\alpha) = Q_k(\alpha)$. In view of (3.3), the function $\Delta g_k(\alpha,p)$ has the property

$$\int_{-\infty}^{\infty} \Delta g_k(\alpha,p)p^j\,dp = 0, \quad j = 0, 1, \dots, k-1, \tag{3.5}$$

and in view of (1.7) and (3.2),

$$\Delta g_k(\alpha,p) = b_k(\alpha)p^{-(k+1)}, \quad |p| > 1. \tag{3.6}$$

Let us now compute $R^{-1}(Y_l P_{kl})$ under the assumptions that P_{kl} is as in (3.2) and (3.3), $l \leq k$, and $l+k$ is even. Denoting $\tilde{P}_{kl} = FP_{kl}$, using equation (3.1) and [12, equation (14.4.48)], we get

$$f_{kl}(r\beta) = \frac{i^l}{(2\pi)^{n/2}} Y_l(\beta) \int_0^\infty \lambda^{n-1} \frac{J_{l+\frac{n-2}{2}}(\lambda r)}{(\lambda r)^{\frac{n-2}{2}}} \tilde{P}_{kl}(\lambda)\,d\lambda. \tag{3.7}$$

Using (3.2), (3.3), and the identity

$$\tilde{P}_{kl}^{(s)}(\lambda) = \frac{\partial^s}{\partial\lambda^s}\left[\frac{1}{(-i\lambda)^j}\int_{-\infty}^{\infty} P_{kl}^{(j)}(p)e^{i\lambda p}\,dp\right], \tag{3.8}$$

where $j > 0$ can be made arbitrarily large, we conclude that

(i) $\tilde{P}_{kl} \in \mathcal{S}(\mathbb{R} \setminus 0)$; that is, $\tilde{P}_{kl} \in C^\infty(\mathbb{R} \setminus 0)$ and all derivatives of $\tilde{P}_{kl}$ decay faster than any power of λ^{-1} as $\lambda \to \infty$;

(ii) $\tilde{P}_{kl}$ can be represented as $\tilde{P}_{kl}(\lambda) = (-i\lambda)^k \tilde{\rho}_{kl}(\lambda)$, where $\tilde{\rho}_{kl} \in \mathcal{S}(\mathbb{R} \setminus 0)$; and

(iii) $\tilde{\rho}_{kl}$ has the following asymptotic representation:

$$\tilde{\rho}_{kl}(\lambda) = b_{kl}\ln\lambda + b_{0kl} + b_{1kl}\lambda + \cdots, \quad \lambda \to 0^+. \tag{3.9}$$

Equation (3.7) can be written as

$$f_{kl}(r\beta)=\frac{i^l}{(2\pi)^{n/2}}\frac{Y_l(\beta)}{r^{\frac{n-2}{2}}}I_{kl}(r), \qquad I_{kl}(r)=\int_0^\infty \lambda^{\frac{n}{2}+k}\tilde{\rho}_{kl}(\lambda)J_{l+\frac{n-2}{2}}(\lambda r)d\lambda.$$

Using (3.9) and the result on the asymptotics of the Hankel transform [1, pp. 231–233], we see that the asymptotic expansion of $I_{kl}(r)$ as $r\to\infty$ consists of the terms

$$\frac{\ln r}{r^{1+\frac{n}{2}+k}}M\left[J_{l+\frac{n-2}{2}},1+\frac{n}{2}+k\right] \quad \text{and} \quad r^{-(1+\frac{n}{2}+k+j)}, \quad j\geq 0,$$

where $M[J_\mu,z]$ denotes the Mellin transform of J_μ evaluated at z. It is well known that

$$M[J_\mu,z]=\frac{2^{z-1}\Gamma\left(\frac{z+\mu}{2}\right)}{\Gamma\left(\frac{\mu-z+2}{2}\right)},$$

and the corresponding strip of analyticity is $-\mu<\mathrm{Re}z<1.5$ [1, p. 414]. According to [1, Lemma 4.3.2], $M[J_\mu,z]$ can be analytically continued into the right half-plane $\mathrm{Re}z\geq 1.5$ as a holomorphic function. According to our assumptions, the ratio $(l-k)/2$ can be only either 0 or a negative integer. Since $|\Gamma(t)|\to\infty$ as t approaches any nonpositive integer, we get

$$M\left[J_{l+\frac{n-2}{2}},1+\frac{n}{2}+k\right]=\lim_{z\to 1+\frac{n}{2}+k}M[J_{l+\frac{n-2}{2}},z]=0, \quad (l-k)/2=0,-1,-2,\dots.$$

Therefore, the term containing $\ln r$ in the expansion of $I_{kl}(r)$ vanishes, and an asymptotic expansion of $f_l(r\beta)$ is of the type of (1.3):

$$f_{kl}(r\beta)=Y_l(\beta)\left(\frac{c_{0kl}}{r^{n+k}}+\frac{c_{1kl}}{r^{n+k+1}}+\cdots\right). \tag{3.10}$$

Let us now compute $R^{-1}(\Delta g_k)$. Denote

$$G(\alpha,p)=\int_{-\infty}^p\frac{(p-t)^{k-1}}{(k-1)!}\Delta g_k(\alpha,t)dt. \tag{3.11}$$

Clearly, $G(\alpha,p)\in C^\infty(Z)$ and $\partial_p^kG(\alpha,p)=\Delta g_k(\alpha,p)$. Using (3.5), we get an equivalent expression for $G(\alpha,p)$:

$$G(\alpha,p)=-\int_p^\infty\frac{(p-t)^{k-1}}{(k-1)!}\Delta g_k(\alpha,t)dt. \tag{3.11$'$}$$

Substituting (3.6) into (3.11) and (3.11$'$), we find that

$$G(\alpha,p)=cb_k(\alpha)p^{-1}, \quad |p|>1, \tag{3.12}$$

for some nonzero constant c. Fourier transforms of Δg_k and G are related by the equation

$$\Delta\tilde{g}_k(\alpha,\lambda)=(-i\lambda)^k\tilde{G}(\alpha,\lambda). \tag{3.13}$$

Using an equation analogous to (3.8), equation (3.12) yields that $\tilde{G}(\alpha,\lambda) \in \mathcal{S}(S^{n-1} \times (\mathbb{R} \setminus 0))$. Moreover, equation (3.12) implies that the asymptotics of $\tilde{G}(\alpha,\lambda)$ as $\lambda \to 0^+$ equals

$$\begin{aligned}\tilde{G}(\alpha,\lambda) &\overset{\text{up to a } C^\infty(Z) \text{ function}}{=} cb_k(\alpha)\int_{|p|\geq 1}\frac{1}{p}e^{i\lambda p}dp = 2cb_k(\alpha)\int_1^\infty \frac{\sin(\lambda p)}{p}dp \\ &= 2cb_k(\alpha)\left(\frac{\pi}{2} - \operatorname{Si}(\lambda)\right), \quad \lambda \to 0^+,\end{aligned} \tag{3.14}$$

where $\operatorname{Si} \in C^\infty(\mathbb{R})$ is the integral sine. Defining $\partial_\lambda^k \tilde{G}(\alpha, 0) := \lim_{\lambda\to 0^+} \partial_\lambda^k \tilde{G}(\alpha,\lambda)$, $k \geq 0$, we get $\tilde{G}(\alpha,\lambda) \in \mathcal{S}(S^{n-1} \times [0,\infty))$. Substituting (3.13) into (3.1) and applying Lemma 1 from section 4 to the resulting integral, we conclude that $\Delta f_k := R^{-1}(\Delta g_k)$ admits an asymptotic expansion

$$\Delta f_k(r\beta) \sim \sum_{l=k}^{\infty} \frac{\Delta\psi_{kl}(\beta)}{r^{n+l}}, \quad r \to \infty, \quad \Delta\psi_{kl} \in C^\infty(S^{n-1}).$$

Together with (3.10), this implies that $f_k := R^{-1}g_k$ admits an asymptotic expansion

$$f_k(r\beta) \sim \sum_{l=k}^{\infty} \frac{\psi_{kl}(\beta)}{r^{n+l}}, \quad r \to \infty, \quad \psi_{kl} \in C^\infty(S^{n-1}); \tag{3.15}$$

hence a formal expansion for $f = R^{-1}g$ is

$$f(r\beta) \sim \sum_{k=m}^{\infty} \frac{1}{r^{n+k}}\left(\sum_{l=m}^{k}\psi_{kl}(\beta)\right), \quad r \to \infty. \tag{3.16}$$

Thus the kth term from the expansion of g contributes to the kth and the following terms of the expansion for f, and this expansion is of the type of (1.3). Also, for each fixed k, only finitely many g_j's, $j = 0, 1, \ldots, k$, contribute to the term f_k, $f_k(r\beta) = r^{-(n+k)}\psi_k(\beta)$, $r \to \infty$, in expansion (1.3).

We have proved that $f = R^{-1}g$ can be formally represented as in (1.3). To see that this is indeed an asymptotic expansion, we have to show that $|f(r\beta) - f_{\Sigma K}(r\beta)| \leq cr^{-(n+K)}$, $r \to \infty$, where $f_{\Sigma K} = R^{-1}g_{\Sigma K}$ and $g_{\Sigma K} := \sum_{k=m}^{K-1} g_k$. Let $M > K > m$ be sufficiently large. Then we can write

$$g(\alpha,p) = g_{\Sigma K}(\alpha,p) + \sum_{k=K}^{M-1} g_k + \eta_M(\alpha,p), \quad p \to \infty.$$

By what was proved above,

$$R^{-1}\left(\sum_{k=K}^{M-1} g_k\right) = O(r^{-(n+K)}), \quad r \to \infty.$$

From the assumptions of the theorem, $\eta_M \in C^\infty(Z)$, η is even, η_M satisfies the first M moment conditions, and η_M admits the asymptotic expansion $\eta_M \sim \sum_{k=M}^{\infty} g_k$. Therefore, $\frac{\partial^j}{\partial p^j}\eta_M(\alpha,p) = O(p^{-(M+j+1)})$, $j \geq 0$. Define $\tilde{\eta}_M(\alpha,\lambda) := F_{p\to\lambda}\eta_M(\alpha,p)$.

Using an equation analogous to (3.8), we have $\tilde\eta_M \in \mathcal{S}(S^{n-1}\times(\mathbb{R}\setminus 0))$. Suppose that $J>0$ is even. Using the identity

$$(1+r^2)^{J/2} f_M(r\beta) = \mathcal{F}^{-1}\big((1-\Delta_\xi)^{J/2}\tilde\eta_M(\alpha,\lambda)\big),$$
$$\xi=\lambda\alpha,\quad f_M = R^{-1}\eta_M,$$

we see that f_M decays sufficiently fast if $\tilde\eta_M$ is sufficiently smooth at the origin. By construction, η_M satisfies the first M moment conditions; therefore, $\tilde\eta_M$ is C^M at the origin. Since M can be made arbitrarily large, f_M can be made to decay as fast as needed. Retaining in the asymptotic expansion of $f_{\Sigma K}$ only the terms of order $\le O(r^{-(n+K-1)})$, we prove the desired assertion.

From (1.6) and (1.7), it follows that $\tilde g(\alpha,\lambda)$, which is used in (3.1), has at worst logarithmic singularity at $\lambda=0$. In a standard fashion, we obtain that $\tilde g(\alpha,\lambda)\in \mathcal{S}(S^{n-1}\times(\mathbb{R}\setminus 0))$. Therefore, $f\in C^\infty(\mathbb{R})$.

Let Q_s be a homogeneous polynomial of degree $s\ge 0$. By assumption, (1.6) can be differentiated with respect to p. Integrating by parts and using assumptions (1)–(3), it is easy to verify the inclusion $g_s(\alpha,p):=Q_s(\alpha)\partial_p^s g(\alpha,p)\in\mathcal{Y}_{m+s}$. Hence $f_s := R^{-1}g_s$ admits an asymptotic expansion of the type of (1.3) with m replaced by $m+s$. Therefore, Lemma 2 from section 4 implies that expansion (1.3) can be differentiated with respect to $x=r\beta$ any number of times.

Finally, we have to show that if f is defined by (3.1), then $Rf=g$ and such an f is unique. We have proved that if f is defined by (3.1), then f is of the type of (1.3). As in the proof of the inclusion $R\mathcal{X}_m\subset\mathcal{Y}_m$, we get that f is absolutely integrable over any $(n-1)$-plane and $Rf\in C^\infty(Z)$. By Lemma 4.5 from [14], $F(Rf)=\mathcal{F}f$ almost everywhere. Since $f:=\mathcal{F}^{-1}Fg$, the equality $Rf=g$ holds almost everywhere. By continuity, $Rf=g$ everywhere. The uniqueness of f follows easily from the injectivity of F and $\mathcal{F}$. The theorem is proved.

4. Auxiliary results.

LEMMA 1. *Let $h\in\mathcal{S}(S^{n-1}\times[0,+\infty))$ and fix $k\ge 0$. Define*

$$f(x)=\frac{1}{(2\pi)^n}\int_{S^{n-1}}\int_0^\infty \lambda^k h(\alpha,\lambda)\exp(-i\lambda\alpha\cdot x)d\lambda d\alpha. \tag{4.1}$$

Then f admits an asymptotic expansion

$$f(r\beta)\sim\sum_{l=k+1}^\infty \frac{\psi_l(\beta)}{r^l},\quad r\to\infty,\quad \psi_l\in C^\infty(S^{n-1}). \tag{4.2}$$

Let $h(\alpha,p)=F_{p\to\lambda}g(\alpha,\lambda)$, where $g\in\mathcal{S}(Z)$ and g is even. Then clearly $h\in\mathcal{S}(Z)$ and $f=Rg$. In this case, equation (4.2) follows from the results in [14]. Nevertheless, Lemma 1 is more general because we consider functions $g(\alpha,p)$ which decay only polynomially as $|p|\to\infty$ (see, e.g., (3.12)). In this case, $h(\alpha,p)=F_{p\to\lambda}g(\alpha,\lambda)$ is not smooth across $\lambda=0$ (the one-sided limits of h and its derivatives exist at $\lambda=0$, however), and hence $h\notin\mathcal{S}(Z)$.

Proof of Lemma 1. Fix $\beta\in S^{n-1}$ and let $x=r\beta$, $r>0$. For a function φ defined on S^{n-1}, we have

$$\int_{S^{n-1}}\varphi(\alpha)d\alpha=\int_{-1}^1\int_{\alpha\cdot\beta=t}\varphi(\alpha)d\alpha\frac{dt}{\sqrt{1-t^2}}$$
$$=\int_{-1}^1\int_{S^{n-2}_{\beta^\perp}}\varphi(t\beta+\sqrt{1-t^2}\,\omega)(1-t^2)^{\frac{n-3}{2}}d\omega\,dt.$$

Therefore, equation (4.1) takes the form

$$f(r\beta) = \int_{-1}^{1}\int_{0}^{\infty} \lambda^k (1-t^2)^{\frac{n-3}{2}} A_\beta(\lambda,t) e^{-i\lambda t r} d\lambda\, dt, \tag{4.3}$$

where

$$A_\beta(\lambda,t) = \frac{1}{(2\pi)^n}\int_{\omega\in S^{n-1}_{\beta^\perp}} h(t\beta+\sqrt{1-t^2}\,\omega,\lambda)d\omega. \tag{4.4}$$

Fix any even $\chi \in C_0^\infty([-1,1])$ such that $\chi(t)\equiv 1$ for $|t|\le 0.5$, and denote $B_\beta(\lambda,t) := (1-t^2)^{\frac{n-3}{2}} A_\beta(\lambda,t)\chi(t)$. Then

$$\begin{aligned} f(r\beta) = &\int_{0.5\le|t|\le 1} (1-t^2)^{\frac{n-3}{2}}(1-\chi(t))\int_0^\infty \lambda^k A_\beta(\lambda,t)e^{-i\lambda t r}d\lambda\, dt \\ &+\int_0^\infty\int_{-1}^{1}\lambda^k B_\beta(\lambda,t)e^{-i\lambda t r}dt\, d\lambda := I_1(r\beta)+I_2(r\beta). \end{aligned} \tag{4.5}$$

Since $h\in \mathcal{S}(S^{n-1}\times[0,+\infty))$, we get that $A_\beta(\lambda,t)$ is $\mathcal{S}([0,\infty))$ in the λ variable and $C([-1,1])$ in the t variable. Therefore, integration by parts gives an asymptotic expansion

$$\int_0^\infty \lambda^k A_\beta(\lambda,t)e^{-i\lambda t r}d\lambda \sim \sum_{l=k+1}^{\infty}\frac{a_l(\beta,t)}{r^l}, \quad 0.5\le|t|\le 1, \quad r\to\infty, \tag{4.6}$$

where

$$a_l(\beta,t) = -(it)^{-l}\frac{\partial^{l-1}}{\partial\lambda^{l-1}}\left(\lambda^k A_\beta(\lambda,t)\right)\bigg|_{\lambda=0}, \quad l\ge k+1.$$

Substituting (4.6) into the definition of I_1 (see (4.5)) and integrating with respect to t over $0.5\le|t|\le 1$, we get

$$\begin{aligned} &I_1(r\beta)\sim\sum_{l=k+1}^{\infty}\frac{\psi_{1l}(\beta)}{r^l}, \\ &\psi_{1l}(\beta) = \int_{0.5\le|t|\le 1}(1-t^2)^{\frac{n-3}{2}}(1-\chi(t))a_l(\beta,t)dt \in C^\infty(S^{n-1}). \end{aligned} \tag{4.7}$$

The inclusion $\psi_{1l}\in C^\infty(S^{n-1})$ holds because $A_\beta(\lambda,t)$ and all of its derivatives with respect to λ depend smoothly on β. (According to (4.4), $A_\beta(\lambda,t)$ is the normalized integral of $h\in\mathcal{S}(S^{n-1}\times[0,\infty))$ over the intersection of the plane $x\cdot\beta=t$ and the unit sphere.) Integration of the asymptotic expansion with respect to the parameter t is justified (cf. Theorem 1.7.5 in [1]).

To find the asymptotic expansion of $I_2(r\beta)$, we transform the integral, denoting $s=\lambda t$ and changing the order of integration:

$$I_2(r\beta) = \int_{-\infty}^{\infty} C_\beta(s)e^{-isr}ds, \qquad C_\beta(s) := \int_{|s|}^{\infty}\lambda^{k-1}B_\beta\left(\lambda,\frac{s}{\lambda}\right)d\lambda. \tag{4.8}$$

Differentiating $C_\beta(s)$ and using the fact that $B_\beta(\lambda,t)\equiv 0$ in neighborhoods of $t=\pm1$, we have $C_\beta^{(m)}(s) = \int_{|s|}^\infty \lambda^{k-1-m}B_\beta^{(m)}(\lambda,s/\lambda)d\lambda$, where $B_\beta^{(m)} := \frac{\partial^m B_\beta}{\partial t^m}$. Clearly, all derivatives of C_β are $C^\infty(\mathbb{R}\setminus 0)$. Moreover, the derivatives $C_\beta^{(m)}(s)$, $m \le k-1$, exist and are continuous at $s=0$. For $m=k$,

$$C_\beta^{(k)}(s) = \int_{|s|}^{\infty} \frac{B_\beta^{(k)}(\lambda, \frac{s}{\lambda})}{\lambda} d\lambda. \tag{4.9}$$

The behavior of $C_\beta^{(k)}(s)$ as $s \to 0$ will be obtained later. Integrating by parts k times in (4.8), we get

$$I_2(r\beta) = \frac{1}{(ir)^k} \int_{-\infty}^{\infty} C_\beta^{(k)}(s) e^{-isr} ds. \tag{4.10}$$

Fix any $\varphi \in C^\infty([0,\infty))$ such that $\varphi(\lambda)=1$ for $0 \le \lambda \le 1/2$ and $\varphi(\lambda)=0$ for $\lambda \ge 1$. Rewrite (4.9) as follows:

$$\begin{aligned} C_\beta^{(k)}(s) = &\int_{|s|}^{\infty} \frac{B_\beta^{(k)}(\lambda, \frac{s}{\lambda}) - B_\beta^{(k)}(0, \frac{s}{\lambda})\varphi(\lambda)}{\lambda} d\lambda \\ &+ \int_{|s|}^{\infty} \frac{B_\beta^{(k)}(0, \frac{s}{\lambda})\varphi(\lambda)}{\lambda} d\lambda =: C_1(s) + C_2(s). \end{aligned} \tag{4.11}$$

Clearly, $C_2 \in C^\infty((0,+\infty))$ and $C_2(s)=0$ for $|s| \ge 1$. Furthermore, for $|s| < 1/2$,

$$\begin{aligned} C_2(s) &= \int_{1/2}^{1} \frac{B_\beta^{(k)}(0, \frac{s}{\lambda})}{\lambda} \varphi(\lambda) d\lambda + \int_{|s|}^{1/2} \frac{B_\beta^{(k)}(0, \frac{s}{\lambda})}{\lambda} d\lambda \\ &= (C^\infty - \text{fn}) + \int_{|s|}^{1/2} \frac{B_\beta^{(k)}(0, \frac{s}{\lambda}) - B_\beta^{(k)}(0,0)}{\lambda} d\lambda + B_\beta^{(k)}(0,0) \int_{|s|}^{1/2} \frac{1}{\lambda} d\lambda \\ &= (C^\infty - \text{fn}) - B_\beta^{(k)}(0,0) \ln|s| + \int_{2|s|}^{1} \frac{B_\beta^{(k)}(0, t\text{sgn}s) - B_\beta^{(k)}(0,0)}{t} dt \\ &= (C^\infty - \text{fn}) - B_\beta^{(k)}(0,0) \ln|s| + \int_{0}^{1} \frac{B_\beta^{(k)}(0, t) - B_\beta^{(k)}(0,0)}{t} dt [\text{sgn}s]^k, \end{aligned} \tag{4.12}$$

where we have used the fact that $B_\beta^{(k)}(0,t)$ is even if k is even and $B_\beta^{(k)}(0,t)$ is odd if k is odd. Thus $C_2(s)$ has the asymptotic expansion of the form

$$\begin{aligned} &C_2(s) \sim b_{-2}(\beta)\ln|s| + b_{-1}(\beta)[\text{sgn}s]^k + \sum_{l=0}^{\infty} b_l(\beta)s^l, \quad s \to 0; \\ &b_l \in C^\infty(S^{n-1}), \quad l \ge -2. \end{aligned}$$

Furthermore, this expansion can be differentiated with respect to s. Using the result from [1, pp. 231–233], we get

$$\int_{-\infty}^{\infty} C_2(s) e^{-isr} ds \sim \sum_{l=1}^{\infty} \frac{c_l(\beta)}{r^l}, \quad c_l \in C^\infty(S^{n-1}). \tag{4.13}$$

Let us denote

$$B_1(\lambda, t) := \frac{B_\beta^{(k)}(\lambda, t) - B_\beta^{(k)}(0,t)\varphi(\lambda)}{\lambda}$$

and represent $C_1(s)$ as

$$C_1(s) = \int_{|s|}^{\infty} B_1(\lambda, s/\lambda) d\lambda.$$

Clearly, $B_1 \in C^\infty([0,\infty) \times \mathbb{R})$ and $B_1(\lambda, t)$ with all derivatives decay faster than any power of λ^{-1} as $\lambda \to \infty$. Furthermore, $B_1(\lambda, t) = 0$ in neighborhoods of $t = \pm 1$, and

$$C_1'(s) = \int_{|s|}^{\infty} \frac{B_1^{(1)}(\lambda, \frac{s}{\lambda})}{\lambda} d\lambda. \tag{4.14}$$

From (4.10) and (4.11), we get

$$\int_{-\infty}^{\infty} C_\beta^{(k)}(s) e^{-isr} ds = \frac{1}{ir} \int_{-\infty}^{\infty} C_1'(s) e^{-isr} ds + \int_{-\infty}^{\infty} C_2(s) e^{-isr} ds. \tag{4.15}$$

Comparing (4.14) with (4.9), we see that the asymptotics of $\int_{-\infty}^{\infty} C_1'(s) e^{-isr} ds$ as $r \to \infty$ can be obtained analogously to (4.11)–(4.12) with $B_\beta^{(k)}$ replaced by $B_1^{(1)}$. Moreover, this integral is multiplied by $1/(ir)$ in (4.15). Therefore, the standard argument that is used for the justification of the integration-by-parts procedure for obtaining asymptotic expansions of integrals (see, e.g., [1, p. 71]) shows that iterating the procedure consisting of (4.11)–(4.15), we obtain the asymptotic expansion of $I_2(r\beta)$, and this expansion is of the form

$$I_2(r\beta) \sim \sum_{l=k+1}^{\infty} \frac{\psi_{2l}(\beta)}{r^l}, \quad \psi_{2l} \in C^\infty(S^{n-1}). \tag{4.16}$$

Combining (4.5), (4.7), and (4.16) proves Lemma 1. □

Lemma 2. *Let $f \in C^\infty(\mathbb{R}^n)$ be such that for any homogeneous polynomial Q_s of degree $s \geq 0$, the function $Q_s(\partial_x) f$ admits the asymptotic expansion*

$$Q_s(\partial_x) f(r\beta) \sim \sum_{k=m+s}^{\infty} \frac{\mu_k(\beta)}{r^k}, \quad r \to \infty, \quad \mu_k \in C^\infty(S^{n-1}), \quad x = r\beta, \tag{4.17}$$

for some $m \geq 1$, where μ_k's depend on Q_s. Then the asymptotic expansion of f can be differentiated with respect to x any number of times. More precisely, if

$$f(r\beta) \sim \sum_{k=m}^{\infty} \frac{\psi_k(\beta)}{r^k}, \quad r \to \infty, \quad \psi_k \in C^\infty(S^{n-1}), \quad x = r\beta, \tag{4.18}$$

then

$$Q_s(\partial_x) f(r\beta) \sim \sum_{k=m}^{\infty} Q_s(\partial_x) \left(\frac{\psi_k(\beta)}{r^k} \right), \quad r \to \infty, \quad x = r\beta. \tag{4.19}$$

Proof. Clearly, it is sufficient to prove (4.19) for only $s = 1$ and $Q_s(\partial_x) = \partial_{x_1}$. By assumption,

$$\partial_{x_1} f(r\beta) \sim \sum_{k=m+1}^{\infty} \frac{\mu_k(\beta)}{r^k} \tag{4.20}$$

for some $\mu_k \in C^\infty(S^{n-1})$. Since $f(x) \to 0$ as $|x| \to \infty$ (see (4.18)), we have

$$-f(x_1, y) = \int_{x_1}^{\infty} \partial_t f(t, y)dt, \quad x = (x_1, y), \quad y \in \mathbb{R}^{n-1}, \tag{4.21}$$

for any $x_1 \in \mathbb{R}$ and $y \in \mathbb{R}^{n-1}$. Substitute expansion (4.20) into (4.21) and integrate formally with respect to t to get

$$\begin{aligned}\int_{x_1}^{\infty} \partial_t f(t, y)dt &\sim \sum_{k=m+1}^{\infty} \int_{x_1}^{\infty} \frac{\mu_k\big(\frac{te_1+y}{(t^2+|y|^2)^{1/2}}\big)}{(t^2+|y|^2)^{k/2}}dt \\ &= \sum_{k=m+1}^{\infty} |y|^{-k+1} \int_{v}^{\infty} \frac{\mu_k\big(\frac{qe_1+y_0}{(q^2+1)^{1/2}}\big)}{(q^2+1)^{k/2}}dq,\end{aligned} \tag{4.22}$$

where e_1 is the unit vector along the x_1-axis, $q = t/|y|$, $y_0 = y/|y|$, and $v = x_1/|y|$. In what follows, we suppose that y_0 and v are fixed. Then the right-hand side of (4.22) has the form of an asymptotic expansion as $|y| \to \infty$. To justify the integration with respect to t, we have to show that if

$$\eta(r\beta) := \partial_{x_1} f(r\beta) - \sum_{k=m+1}^{K-1} \frac{\mu_k(\beta)}{r^k} = O(r^{-K}),$$

then $\int_{x_1}^{\infty} \eta(t, y)dt = O(|y|^{-K+1})$. Since $\eta(r\beta) = O(r^{-K})$, we get $|\eta(r\beta)| \leq cr^{-K}$, $r \geq 1$, for some $c > 0$. Using this estimate and integrating with respect to t analogously to (4.22), we obtain the desired assertion.

Using the variables y_0 and v, transform (4.18):

$$f(x_1, y) \sim \sum_{k=m}^{\infty} |y|^{-k} \frac{\psi_k\big(\frac{ve_1+y_0}{(v^2+1)^{1/2}}\big)}{(v^2+1)^{k/2}}. \tag{4.23}$$

Since the asymptotic expansion—if it exists—is unique, we conclude from (4.21)–(4.23) by taking $|y| \to \infty$ that

$$-\frac{\psi_k\big(\frac{ve_1+y_0}{(v^2+1)^{1/2}}\big)}{(v^2+1)^{k/2}} = \int_{v}^{\infty} \frac{\mu_{k+1}\big(\frac{qe_1+y_0}{(q^2+1)^{1/2}}\big)}{(q^2+1)^{(k+1)/2}}dq, \quad k \geq m, \quad v \in \mathbb{R}, \quad y_0 \in S^{n-2}.$$

Therefore,

$$\frac{\partial}{\partial v}\left(\frac{\psi_k\big(\frac{ve_1+y_0}{(v^2+1)^{1/2}}\big)}{(v^2+1)^{k/2}}\right) = \frac{\mu_{k+1}\big(\frac{ve_1+y_0}{(v^2+1)^{1/2}}\big)}{(v^2+1)^{(k+1)/2}}, \quad k \geq m, \quad v \in \mathbb{R}, \quad y_0 \in S^{n-2}.$$

Multiplying the last equation by $|y|^{-k}$ on both sides, returning to the variables x_1 and y, and then denoting $r\beta = x_1e_1 + y$, we get

$$\frac{\partial}{\partial x_1}\left(\frac{\psi_k(\beta)}{r^k}\right) = \frac{\mu_{k+1}(\beta)}{r^{k+1}}.$$

Lemma 2 is proved. □

Acknowledgment. The author thanks Professor Alexander G. Ramm for useful discussions.

REFERENCES

[1] N. Bleistein and R. Handelsman, *Asymptotic Expansions of Integrals*, Dover, Mineola, NY, 1986.

[2] I. Gelfand, M. Graev, and N. Vilenkin, *Integral Geometry and Representation Theory*, Academic Press, New York, 1965.

[3] S. Helgason, *The Radon transform on euclidean spaces, compact two-point homogeneous spaces and Grassmann manifolds*, Acta Math., 113 (1965), pp. 153–179.

[4] S. Helgason, *The Radon Transform*, Birkhauser, Boston, 1980.

[5] S. Helgason, *Ranges of Radon transform*, in Proc. Symposia in Applied Mathematics, Vol. 27, L. Shepp, ed., AMS, Providence, RI, 1982, pp. 63–70.

[6] A. Hertle, *Continuity of the Radon transform and its inverse on Euclidean spaces*, Math. Z., 184 (1983), pp. 165–192.

[7] A. Hertle, *On the range of Radon transform and its dual*, Math. Ann., 267 (1984), pp. 91–99.

[8] D. Ludwig, *The Radon transform on Euclidean spaces*, Comm. Pure Appl. Math., 19 (1966), pp. 49–81.

[9] P. Lax and R. Phillips, *The Paley–Wiener theorem for the Radon transform*, Comm. Pure Appl. Math., 23 (1970), pp. 409–424.

[10] A. Louis, *Orthogonal function series expansions and the null space of the Radon transform*, SIAM J. Math. Anal., 15 (1984), pp. 621–633.

[11] A. G. Ramm, *The Radon transform is an isomorphism between $L^2(B)$ and $H_e(Z_a)$*, Appl. Math. Lett., 8 (1995), pp. 25–29.

[12] A. G. Ramm and A. I. Katsevich, *The Radon Transform and Local Tomography*, CRC Press, Boca Raton, FL, 1996.

[13] K. Smith, D. Solmon, and S. Wagner, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. Amer. Math. Soc., 83 (1977), pp. 1227–1270.

[14] D. Solmon, *Asymptotic formulas for the dual Radon transform and applications*, Math. Z., 195 (1987), pp. 321–343.

[15] D. Solmon and W. Madych, *A range theorem for the Radon transform*, Proc. Amer. Math. Soc., 104 (1988), pp. 79–85.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 867–885, July 1997

NONMINIMIZING POSITIVE SOLUTIONS FOR EQUATIONS WITH CRITICAL EXPONENTS IN THE HALF-SPACE*

GIOVANNA CERAMI† AND DONATO PASSASEO‡

Abstract. This paper is concerned with the existence of positive solutions of the nonlinear elliptic problem $-\Delta u + a(x)u = u^{(N+2)/(N-2)}$, $a(x) \geq 0$, with Neumann boundary conditions in a half-space $\Pi \subset \mathbb{R}^N$, $N \geq 3$. The main feature of the problem is a "double" lack of compactness due to the unboundedness of the domain and the presence of the critical Sobolev exponent. The solutions are searched using variational methods, although the functional related to the problem does not satisfy the Palais–Smale compactness condition. We observe that the problem considered has no solutions if $a(x)$ is a positive constant; conditions on $a(x)$ are given sufficient to guarantee existence and multiplicity of positive solutions.

Key words. nonlinear elliptic equations, critical Sobolev exponent, positive solutions

AMS subject classifications. 35J65, 35J20

PII. S0036141095295747

1. Introduction and statement of the results.

This paper deals with the following problem: find solutions of finite energy of

$$\text{(P)} \qquad \begin{cases} -\Delta u + a(x)u = u^p & \text{in } \mathbb{R}^N_+, \\ u > 0 & \text{in } \mathbb{R}^N_+, \\ \partial u/\partial x_N = 0 & \text{on } \partial\mathbb{R}^N_+, \\ u(x) \to 0 & \text{as } |x| \to +\infty, \end{cases}$$

where $p = 2^* - 1$, $2^* = 2N/(N-2)$, $N \geq 3$, $a(x) \geq 0$, $\mathbb{R}^N_+ = \{x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N : x_N > 0\}$, and $\partial\mathbb{R}^N_+ = \{x = (x_1, x_2, \ldots, x_{N-1}, 0) \in \mathbb{R}^N\}$.

If $a(x) = 0$, (P) has the positive solution

$$U(x) = \frac{[N(N-2)]^{(N-2)/4}}{(1+|x|^2)^{(N-2)/2}},$$

and all of the positive solutions can be obtained from this one by translations and scale changes, namely, they have the form $\sigma^{-(N-2)/4}U((x-y)/\sqrt{\sigma})$, $\sigma > 0$, $y \in \partial\mathbb{R}^N_+$. On the other hand, if $a(x) = \lambda > 0$ (λ constant), it is not difficult to realize that (P) has no solutions. In fact, if u were a solution of (P), its extension to all of $\mathbb{R}^N$, obtained by reflection, would solve $-\Delta u + \lambda u = u^{2^*-1}$ in $\mathbb{R}^N$, and this problem has no nontrivial solutions, which follows from a generalized version of the Pohozaev identity (see, for instance, [BL] or [BC]).

On the other hand, it has been shown (see [BC], [P]) that the presence of a nonconstant coefficient $a(x)$ plays an important role in finding finite-energy positive solutions of the following equation: $-\Delta u + a(x)u = u^{2^*-1}$ in $\mathbb{R}^N$.

*Received by the editors December 4, 1995; accepted for publication March 12, 1996. This research was supported by Ministero della Universitá della Ricerca Scientifica e Tecnologica (MURST) research funds. The authors are members of the Consiglio Nazionale della Ricerche (CNR)/Gruppo Nazionale per l'Analisi Funzionale e le sue Applicazioni (GNAFA).

http://www.siam.org/journals/sima/28-4/29574.html

†Facoltà di Ingegneria, Università di Palermo, 90100 Palermo, Italy (cerami@math.unipa.it).

‡Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56127 Pisa, Italy (passaseo@dm.unipi.it).

In this paper, we consider functions $a(x)$ that satisfy the following assumption:

$$\tag{1.1} \begin{cases} \text{(i)} \quad \lim_{|x|\to+\infty} a(x) = a_\infty \geq 0, & a_\infty \in \mathbb{R}, \\ \text{(ii)} \quad a(x) \geq a_\infty & \forall x \in \mathbb{R}^N_+, \\ \text{(iii)} \quad a(x) - a_\infty \in L^{N/2}(\mathbb{R}^N_+), & |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)} \neq 0. \end{cases}$$

The Hilbert spaces that we obtain as the closure of $\mathcal{C}^\infty(\mathbb{R}^N_+)$ with respect to

$$\|u\|_{\mathcal{D}} = \left(\int_{\mathbb{R}^N_+} |\nabla u|^2 dx\right)^{1/2}, \qquad \|u\|_W = \left[\int_{\mathbb{R}^N_+} (|\nabla u|^2 + u^2) dx\right]^{1/2}$$

are denoted $\mathcal{D}^{1,2}(\mathbb{R}^N_+)$ and $W^{1,2}(\mathbb{R}^N_+)$, respectively. The results that we obtain can be stated as follows.

THEOREM 1.1. *Let $a(x)$ satisfy* (1.1) *and let $a_\infty > 0$. Then there exists a positive number $\mathcal{A}$ such that if $a_\infty \in (0, \mathcal{A})$, then* (P) *admits at least a positive solution $v \in W^{1,2}(\mathbb{R}^N_+)$. Furthermore, if the condition*

$$\tag{1.2} |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)} < (1 - 2^{-2/N})S$$

is satisfied, (P) *has at least another solution $u \in W^{1,2}(\mathbb{R}^N_+)$.*

THEOREM 1.2. *Let $a(x)$ satisfy* (1.1) *and let $a_\infty = 0$. Assume that* (1.2) *holds, i.e.,*

$$|a|_{L^{N/2}(\mathbb{R}^N_+)} < (1 - 2^{-2/N})S;$$

then (P) *has at least one solution $u \in \mathcal{D}^{1,2}(\mathbb{R}^N_+)$.*

Problem (P) has a variational structure, so although the lack of compactness (due to the unboundedness of $\mathbb{R}^N_+$ and the presence of the critical Sobolev exponent) gives rise to some difficulties, the investigation is carried out using variational methods. The solutions of (P) correspond to the positive functions that are critical points of the energy functional

$$\tag{1.3} E(u) = \int_{\mathbb{R}^N_+} (|\nabla u|^2 + a(x)u^2) dx$$

constrained on the manifold

$$\tag{1.4} V = \left\{ u \in H : \int_{\mathbb{R}^N_+} |u|^{2^*} dx = 1 \right\},$$

where H is either $W^{1,2}(\mathbb{R}^N_+)$ or $\mathcal{D}^{1,2}(\mathbb{R}^N_+)$ according to whether a_∞ is positive or equal to zero.

In both cases, (P) cannot be solved by minimization. In fact, we shall see in section 2 that E does not achieve its infimum Σ on V. Therefore, the study of the problem needs more subtle tools, such as minimax theory. Clearly, all of the solutions whose existence is stated in Theorems 1.1 and 1.2 correspond to positive functions for which $E(u) > \Sigma$. However, it is worth mentioning their different natures. Indeed, the proofs of these theorems will make clear that the energy of the "first" solution, v, found in the case $a_\infty > 0$, is very near to Σ and that this solution vanishes when

$a_\infty \to 0$. On the other hand, the solutions u are of high energy, namely, their energy is bounded from below by a number that is independent of a_∞ and strictly larger than Σ.

Finally, let us point out that an easy scale change shows that to any solution of (P) with $a(x) = a_\infty + \alpha(x)$ $(\alpha(x) = a(x) - a_\infty)$ there corresponds a solution of (P) with $a(x) = \lambda a_\infty + \lambda\alpha(\sqrt{\lambda}x)$ $\forall\lambda > 0$. Thus the condition on the size of $\lim_{|x|\to+\infty} a(x)$ that appears in Theorem 1.1 also can be expressed by saying that a_∞ can be arbitrarily large, provided that $a(x) - a_\infty$ is a function that is "concentrated" enough. Also, the claims of Theorems 1.1 and 1.2 hold when $\mathbb{R}^N_+$ in (P) is replaced by any half-space Π with the boundary condition $\partial u/\partial\nu = 0$, where ν is the outer normal to $\partial\Pi$.

This paper is organized as follows: In section 2, some useful facts are recalled, a nonexistence theorem is proven, the compactness question is discussed, and some basic estimates are stated. Section 3 contains the proofs of the theorems.

2. Preliminary remarks, some useful facts, and estimates. We begin by recalling some definitions and known facts.

S denotes the best Sobolev constant, i.e.,

$$S = \inf\left\{\int_{\mathbb{R}^N} |\nabla u|^2 dx : u \in \mathcal{D}^{1,2}(\mathbb{R}^N),\ |u|_{L^{2^*}(\mathbb{R}^N)} = 1\right\}. \tag{2.1}$$

S is achieved by the function

$$\Psi_{1,0}(x) = \bar{U}(x)/|\bar{U}|_{L^{2^*}(\mathbb{R}^N)}, \quad \text{where } \bar{U} = \frac{1}{[1+|x|^2]^{\frac{N-2}{2}}},$$

and all of the minimizers for S are the functions

$$\begin{gathered}\Psi_{\sigma,y} = \sigma^{-\frac{N-2}{4}}\Psi_{1,0}\left(\frac{x-y}{\sqrt{\sigma}}\right) = \frac{1}{|\bar{U}|_{L^{2^*}(\mathbb{R}^N)}}\frac{\sigma^{(N-2)/4}}{[\sigma+|x-y|^2]^{\frac{N-2}{2}}}, \\ \sigma \in R^+\backslash\{0\}, \qquad y \in \mathbb{R}^N,\end{gathered} \tag{2.2}$$

obtained from $\Psi_{1,0}$ by translation and rescaling.

We set

$$\Sigma = \inf\left\{\int_{\mathbb{R}^N_+} |\nabla u|^2 dx : u \in \mathcal{D}^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1\right\}. \tag{2.3}$$

If we consider the definition and properties of S, it is not difficult to verify that

$$\Sigma = 2^{-2/N}S,$$

that Σ is achieved by the function defined $\forall x \in \mathbb{R}^N_+$ by

$$\tilde{\Psi}_{1,0}(x) = 2^{1/2^*}\Psi_{1,0}(x),$$

and that all of the minimizers for Σ are the functions $\tilde{\Psi}_{\sigma,y}$, $\sigma \in \mathbb{R}^+\backslash\{0\}$, $y \in \partial\mathbb{R}^N_+$, defined by

$$\tilde{\Psi}_{\sigma,y}(x) = \sigma^{-\frac{(N-2)}{4}}\tilde{\Psi}_{1,0}\left(\frac{x-y}{\sqrt{\sigma}}\right). \tag{2.4}$$

We remark that the infima in (2.1) and (2.3) do not change if we restrict our consideration to the functions u that belong to $W^{1,2}(\mathbb{R}^N)$ and $W^{1,2}(\mathbb{R}^N_+)$, respectively.

Let us prove now a nonexistence result.

Proposition 2.1. *Let a satisfy* (1.1). *Set*

$$\inf\{E(u):\ u \in V\} = \Sigma_a. \tag{2.5}$$

Then

$$\Sigma_a = \Sigma$$

and the minimization problem (2.5) *has no solution.*

Proof. Since $a(x) \geq 0$ in $\mathbb{R}^N_+$, obviously $\Sigma_a \geq \Sigma$. To show that the equality holds, let us consider the sequence

$$\Phi_{\frac{1}{n},0}(x) = \chi(|x|)\tilde{\Psi}_{\frac{1}{n},0}(x),$$

where $\chi \in \mathcal{C}^\infty([0,+\infty))$ is a nonincreasing real function such that $\chi(t) = 1$ if $t \in [0,1/2]$ and $\chi(t) = 0$ if $t \geq 1$. Well-known computations (see, for example, [BN]) give

$$\frac{\int_{\mathbb{R}^N_+}[|\nabla\Phi_{\frac{1}{n},0}(x)|^2 + a_\infty\Phi^2_{\frac{1}{n},0}(x)]dx}{|\Phi_{\frac{1}{n},0}(x)|^2_{L^{2^*}(\mathbb{R}^N_+)}} = \begin{cases} \Sigma + O\left(\frac{1}{n}\right), & N \geq 5, \\ \Sigma + O\left(\frac{1}{n}|\log\frac{1}{n}|\right), & N = 4, \\ \Sigma + O\left(1/\sqrt{n}\right), & N = 3. \end{cases} \tag{2.6}$$

On the other hand, setting $\alpha(x) = a(x) - a_\infty$ and $B_\rho(0) = \{x \in \mathbb{R}^N : |x| < \rho\}$, we have $\forall\rho > 0$

$$\begin{aligned}\int_{\mathbb{R}^N_+}\alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx &= \int_{\mathbb{R}^N_+\cap B_\rho(0)}\alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx + \int_{\mathbb{R}^N_+\setminus B_\rho(0)}\alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx \\ &\leq |\Phi_{\frac{1}{n},0}(x)|^2_{L^{2^*}(\mathbb{R}^N_+)} \cdot \left[\int_{\mathbb{R}^N_+\cap B_\rho(0)}(\alpha(x))^{N/2}dx\right]^{2/N} \\ &\quad + |\alpha|_{L^{N/2}(\mathbb{R}^N_+)}\left[\int_{\mathbb{R}^N_+\setminus B_\rho(0)}\Phi^{2^*}_{\frac{1}{n},0}(x)dx\right]^{2/2^*},\end{aligned}$$

so

$$\begin{aligned}&\frac{\int_{\mathbb{R}^N_+}\alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx}{|\Phi_{\frac{1}{n},0}|^2_{L^{2^*}(\mathbb{R}^N_+)}} \\ &\leq \left[\int_{\mathbb{R}^N_+\cap B_\rho(0)}(\alpha(x))^{N/2}dx\right]^{2/N} + \frac{|\alpha|_{L^{N/2}(\mathbb{R}^N_+)}\left[\int_{\mathbb{R}^N_+\setminus B_\rho(0)}\Phi^{2^*}_{\frac{1}{n},0}(x)dx\right]^{2/2^*}}{|\Phi_{\frac{1}{n},0}|^2_{L^{2^*}(\mathbb{R}^N_+)}}.\end{aligned}$$

Now

$$\lim_{n\to+\infty}\int_{\mathbb{R}^N_+\setminus B_\rho(0)}\Phi^{2^*}_{\frac{1}{n},0}(x)dx = 0$$

and

$$\lim_{n\to+\infty}|\Phi_{\frac{1}{n},0}|_{L^{2^*}(\mathbb{R}^N_+)} = 1.$$

Hence $\forall \rho > 0$,

$$\lim_{n\to+\infty} \frac{\int_{\mathbb{R}^N_+} \alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx}{|\Phi_{\frac{1}{n},0}|^2_{L^{2^*}(\mathbb{R}^N_+)}} \leq \left[\int_{\mathbb{R}^N_+\cap B_\rho(0)} (\alpha(x))^{N/2}dx\right]^{2/N}.$$

Thus from $\alpha \in L^{N/2}(\mathbb{R}^N_+)$, we deduce

$$\lim_{\rho\to 0} \left[\int_{\mathbb{R}^N_+\cap B_\rho(0)} (\alpha(x))^{N/2}dx\right]^{2/N} = 0$$

and then

$$\lim_{n\to+\infty} \frac{\int_{\mathbb{R}^N_+} \alpha(x)\Phi^2_{\frac{1}{n},0}(x)dx}{|\Phi_{\frac{1}{n},0}|^2_{L^{2^*}(\mathbb{R}^N_+)}} = 0. \tag{2.7}$$

Therefore, equations (2.6) and (2.7) give $\lim_{n\to+\infty} E(\Phi_{1/n,0}(x)/|\Phi_{1/n,0}|_{L^{2^*}(\mathbb{R}^N_+)}) = \Sigma$, as desired. □

Let us now assume that the minimization problem (2.5) has a solution u and—without loss of generality—that $u \geq 0$. Let us denote by u^* and a^* the extensions by reflection to all of $\mathbb{R}^N$ of u and a, respectively. Then

$$\frac{\int_{\mathbb{R}^N}(|\nabla u^*|^2 + a^*(x)(u^*(x))^2)dx}{|u^*|^2_{L^{2^*}(\mathbb{R}^N)}} = S,$$

so we have

$$S \leq \frac{\int_{\mathbb{R}^N}|\nabla u^*|^2 dx}{|u^*|^2_{L^{2^*}(\mathbb{R}^N)}} \leq \frac{\int_{\mathbb{R}^N}(|\nabla u^*|^2 + a^*(x)(u^*(x))^2)dx}{|u^*|^2_{L^{2^*}(\mathbb{R}^N)}} = S.$$

The above relation implies that $\int_{\mathbb{R}^N} a^*(x)(u^*(x))^2dx = 0$ and $u^* = \Psi_{\sigma,y}$ for some $\sigma > 0$ and $y \in \mathbb{R}^N$. Thus, using the assumptions on a and the fact that $\Psi_{\sigma,y}(x) > 0$ $\forall x \in \mathbb{R}^N$, we deduce

$$0 = \int_{\mathbb{R}^N} a^*(x)(u^*(x))^2dx = \int_{\mathbb{R}^N} a^*(x)\Psi^2_{\sigma,y}(x)dx > 0,$$

which is impossible.

The following lemma states a lower bound for the energy of a sign-changing critical point u of E on V.

LEMMA 2.2 *Let a satisfy* (1.1). *Let u be a critical point of E on V. If $E(u) < S$, then the function u does not change sign.*

Proof. We argue by contradiction and assume that $u = u^+ + u^-$, $u^+ \neq 0$, and $u^- \neq 0$. By Proposition 2.1,

$$\Sigma|u^\pm|^2_{L^{2^*}(\mathbb{R}^N_+)} < \int_{\mathbb{R}^N_+}(|\nabla u^\pm|^2 + a(x)(u^\pm)^2)dx$$

and, since u is a critical point of E on V,

$$\int_{\mathbb{R}^N_+}[|\nabla u^\pm|^2 + a(x)(u^\pm)^2]dx = E(u)\int_{\mathbb{R}^N_+}|u^\pm|^{2^*}dx.$$

Then

$$|u^{\pm}|^{2^*}_{L^{2^*}(\mathbb{R}^N_+)} \geq \left(\frac{\Sigma}{E(u)}\right)^{N/2},$$

which, considering that $|u|_{L^{2^*}(\mathbb{R}^N_+)} = 1$, gives

$$E(u) \geq 2^{2/N}\Sigma = S,$$

which contradicts our assumption. □

The following proposition gives useful information about the compactness of E on V.

Proposition 2.3. *Let a satisfy* (1.1). *Then the pair* (E, V) *verifies the Palais–Smale condition in the energy range* (Σ, S), *i.e., any sequence* $\{u_n\}$ *such that*

$$\begin{cases} u_n \in V, \quad \operatorname{grad} E|_V(u_n) \to 0 \qquad \textit{in } H^*, \\ \lim\limits_{n\to+\infty} E(u_n) = c \in (\Sigma, S) \end{cases} \tag{2.8}$$

is relatively compact.

Proof. Let u_n be a sequence of functions that satisfy (2.8). Let us denote by u_n^* and a^* the functions obtained by u_n and a extending to all of $\mathbb{R}^N$ by reflection, respectively. Then we have $u_n^* \in W^{1,2}(\mathbb{R}^N)$ if $a_\infty > 0$, $u_n^* \in \mathcal{D}^{1,2}(\mathbb{R}^N)$ if $a_\infty = 0$, and

$$\left|\frac{u_n^*}{2^{1/2^*}}\right|_{L^{2^*}(\mathbb{R}^N)} = 1, \qquad \frac{1}{2^{2/2^*}}\int_{\mathbb{R}^N}[|\nabla u_n^*|^2 + a^*(x)(u_n^*(x))^2]dx \to 2^{2/N}c,$$

$$\int_{\mathbb{R}^N}[(\nabla u_n^*|\nabla v) + a^*(x)u_n^*v]dx + (2^{2/N}c + o(1))\int_{\mathbb{R}^N}|u_n^*|^{2^*-2}u_n^*v\,dx = o(1)$$

$\forall v \in W^{1,2}(\mathbb{R}^N)$ (respectively, $v \in \mathcal{D}^{1,2}(\mathbb{R}^N)$ if $a_\infty = 0$).

Since $2^{2/N}c \in (S, 2^{2/N}S)$, by Theorem 2.5 and Corollary 2.10 of [BC] and Lemma 1.9 of [P], u_n^* is relatively compact, i.e., converges strongly, up to a subsequence, to a function u^* that verifies $u^*(x_1, x_2, \dots, x_N) = u^*(x_1, x_2, \dots, -x_N)$, and this yields the desired result. □

Let us denote by π the projection on $\partial\mathbb{R}^N_+$, i.e.,

$$\pi : \mathbb{R}^N \to \partial\mathbb{R}^N_+, \quad \pi(x_1, x_2, \dots, x_N) = (x_1, x_2, \dots, x_{N-1}, 0),$$

and $\forall\rho > 0$, set

$$\Lambda_\rho(y) = \{x \in \mathbb{R}^N_+ : |\pi(x) - \pi(y)| < \rho\}.$$

Let us define the maps $\beta : H \to \partial\mathbb{R}^N_+$ and $\gamma : H \to \mathbb{R}$ by

$$\beta(u) = \int_{\mathbb{R}^N_+}\frac{\pi(x)}{1 + |\pi(x)|}|u(x)|^{2^*}dx \Big/ |u|^{2^*}_{L^{2^*}(\mathbb{R}^N_+)},$$

$$\gamma(u) = \int_{\mathbb{R}^N_+}\left|\frac{\pi(x)}{1 + |\pi(x)|} - \beta(u)\right||u(x)|^{2^*}dx \Big/ |u|^{2^*}_{L^{2^*}(\mathbb{R}^N_+)}.$$

The following two propositions provide useful estimates.

PROPOSITION 2.4. *Let $\alpha(x) \in L^{N/2}(\mathbb{R}^N_+)$ be a nonegative function such that $|\alpha|_{L^{N/2}(\mathbb{R}^N_+)} \neq 0$. Then*

$$\inf\left\{ \int_{\mathbb{R}^N_+} [|\nabla u|^2 + \alpha(x)u^2]dx : u \in \mathcal{D}^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1, \right.$$
$$\left. \beta(u) = 0,\ \gamma(u) = \frac{1}{3} \right\} > \Sigma. \tag{2.9}$$

Proof. Clearly,

$$\inf\left\{ \int_{\mathbb{R}^N_+} [|\nabla u|^2 + \alpha(x)u^2]dx : u \in \mathcal{D}^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1, \right.$$
$$\left. \beta(u) = 0,\ \gamma(u) = \frac{1}{3} \right\} \geq \Sigma.$$

To prove (2.9), we argue by contradiction and suppose that equality holds in the above relation. Thus we can find a sequence $\{u_n\}$ such that $u_n \in \mathcal{D}^{1,2}(\mathbb{R}^N_+)$ and

$$\begin{cases} \text{(a)} \quad |u_n|_{L^{2^*}(\mathbb{R}^N_+)} = 1, \qquad \beta(u_n) = 0, \qquad \gamma(u_n) = \frac{1}{3}, \\ \text{(b)} \quad \lim_{n\to+\infty} \int_{\mathbb{R}^N_+} [|\nabla u_n|^2 + \alpha(x)u_n^2)]dx = \Sigma. \end{cases} \tag{2.10}$$

Therefore, since $\alpha(x) \geq 0$, from

$$\Sigma = \lim_{n\to+\infty} \int_{\mathbb{R}^N_+} [|\nabla u_n|^2 + \alpha(x)u_n^2)]dx \geq \lim_{n\to+\infty} \int_{\mathbb{R}^N_+} |\nabla u_n|^2 dx \geq \Sigma,$$

we obtain $\lim_{n\to+\infty} \int_{\mathbb{R}^N_+} |\nabla u_n|^2 dx = \Sigma$.

Hence by the uniqueness of the family of functions $\tilde{\Psi}_{\sigma,y}$ (defined in (2.4)) that realize Σ, we deduce that

$$u_n(x) = \tilde{\Psi}_{\sigma_n,y_n}(x) + w_n(x) \quad \forall x \in \mathbb{R}^N_+,$$

where $\sigma_n \in \mathbb{R}^+\backslash\{0\}$, $y_n \in \partial\mathbb{R}^N_+$, and $\{w_n\}$ is a sequence that goes strongly to zero in $\mathcal{D}^{1,2}(\mathbb{R}^N_+)$ and $L^{2^*}(\mathbb{R}^N_+)$.

We claim that, up to subsequences,

$$\text{(a)} \quad \lim_{n\to+\infty} \sigma_n = \bar{\sigma} > 0, \qquad \text{(b)} \quad \lim_{n\to+\infty} y_n = \bar{y} \in \partial\mathbb{R}^N_+. \tag{2.11}$$

Indeed, once (2.11) is shown to be true, the proof can be achieved quickly. In fact, it suffices to observe that in this case $\tilde{\Psi}_{\sigma_n,y_n} \to \tilde{\Psi}_{\bar{\sigma},\bar{y}}$ strongly in $\mathcal{D}^{1,2}(\mathbb{R}^N_+)$ and $L^{2^*}(\mathbb{R}^N_+)$, so from (2.10)(b) it follows that

$$\Sigma = \lim_{n\to+\infty} \int_{\mathbb{R}^N_+} [|\nabla u_n|^2 + \alpha(x)u_n^2)]dx = \lim_{n\to+\infty} \int_{\mathbb{R}^N_+} [|\nabla \tilde{\Psi}_{\sigma_n,y_n}(x)|^2 + \alpha(x)\tilde{\Psi}^2_{\sigma_n,y_n}(x)]dx$$
$$= \int_{\mathbb{R}^N_+} [|\nabla \tilde{\Psi}_{\bar{\sigma},\bar{y}}|^2 + \alpha(x)\tilde{\Psi}^2_{\bar{\sigma},\bar{y}}(x)]dx = \Sigma + \int_{\mathbb{R}^N_+} \alpha(x)\tilde{\Psi}^2_{\bar{\sigma},\bar{y}}(x)dx,$$

which, because of the assumptions on α and the positivity of $\tilde{\Psi}_{\bar{\sigma},\bar{y}}$, is impossible.

Let us now prove the claim in (2.11). To prove (2.11)(a), let us first show that $\{\sigma_n\}$ is bounded. In fact, if for some subsequence (still denoted by σ_n) $\lim_{n\to+\infty}\sigma_n = +\infty$ occurs, then $\forall\rho > 0$,

$$\begin{aligned}\lim_{n\to+\infty}\int_{\Lambda_\rho(0)} |u_n|^{2^*}dx &= \lim_{n\to+\infty}\int_{\Lambda_\rho(0)} \left|\tilde{\Psi}_{\sigma_n,y_n}(x)\right|^{2^*} dx\\ &= \lim_{n\to+\infty}\int_{\Lambda_{\frac{\rho}{\sigma_n}}(0)} \left|\tilde{\Psi}_{1,0}\left(x-\frac{y_n}{\sigma_n}\right)\right|^{2^*} dx = 0.\end{aligned}$$

Then if we consider that $\beta(u_n) = 0$, we have $\forall\rho > 0$

$$\begin{aligned}\gamma(u_n) = \int_{\mathbb{R}^N_+}\frac{|\pi(x)|}{1+|\pi(x)|}|u_n(x)|^{2^*}dx &\geq \int_{\mathbb{R}^N_+\setminus\Lambda_\rho(0)}\frac{|\pi(x)|}{1+|\pi(x)|}|u_n(x)|^{2^*}dx\\ \geq \frac{\rho}{1+\rho}\int_{\mathbb{R}^N_+\setminus\Lambda_\rho(0)}|u_n(x)|^{2^*}dx,\end{aligned}$$

so

$$\liminf_{n\to+\infty}\gamma(u_n) \geq \rho/(1+\rho) \quad \forall\rho > 0,$$

which implies

$$\lim_{n\to+\infty}\gamma(u_n) = 1,$$

which contradicts (2.10)(a).

Thus up to a subsequence, $\lim_{n\to+\infty}\sigma_n = \bar{\sigma}\in\mathbb{R}^+$. If $\bar{\sigma} = 0$ occurs, then $\forall\rho > 0$,

$$\begin{aligned}\lim_{n\to+\infty}\int_{\mathbb{R}^N_+\setminus\Lambda_\rho(y_n)} |u_n|^{2^*}dx &= \lim_{n\to+\infty}\int_{\mathbb{R}^N_+\setminus\Lambda_\rho(y_n)} \left|\tilde{\Psi}_{\sigma_n,y_n}(x)\right|^{2^*} dx\\ &= \lim_{n\to+\infty}\int_{\mathbb{R}^N_+\setminus\Lambda_{\frac{\rho}{\sigma_n}}(\frac{y_n}{\sigma_n})} \left|\tilde{\Psi}_{1,0}\left(x-\frac{y_n}{\sigma_n}\right)\right|^{2^*} dx = 0.\end{aligned}$$

Hence $\forall\rho > 0$,

$$\begin{aligned}\frac{|y_n|}{1+|y_n|} &= \left|\int_{\mathbb{R}^N_+}\left(\frac{y_n}{1+|y_n|}-\frac{\pi}{1+|\pi(x)|}\right)|u_n(x)|^{2^*}dx\right|\\ &\leq \int_{\mathbb{R}^N_+\setminus\Lambda_\rho(y_n)}\left|\frac{y_n}{1+|y_n|}-\frac{\pi}{1+|\pi(x)|}\right||u_n(x)|^{2^*}dx\\ &\quad + \int_{\Lambda_\rho(y_n)}\left|\frac{y_n}{1+|y_n|}-\frac{\pi}{1+|\pi(x)|}\right||u_n(x)|^{2^*}dx \leq \rho + o(1),\end{aligned}$$

which implies $|y_n|\to 0$ as $n\to+\infty$. Now we obtain

$$\begin{aligned}\lim_{n\to+\infty}\gamma(u_n) &= \lim_{n\to+\infty}\int_{\mathbb{R}^N_+}\left|\frac{\pi(x)}{1+|\pi(x)|}-\beta(u_n)\right||u_n(x)|^{2^*}dx\\ &= \lim_{n\to+\infty}\int_{\mathbb{R}^N_+}\left|\frac{\pi(x)}{1+|\pi(x)|}-\frac{y_n}{1+|y_n|}\right||u_n(x)|^{2^*}dx = 0,\end{aligned}$$

which contradicts (2.10)(a). Thus the first part of the claim is proved.

Let us now show that $\{|y_n|\}$ is bounded. We argue by contradiction and suppose that a subsequence $\{y_m\}$ exists for which $\lim_{m\to+\infty}|y_m| = +\infty$. Then $\forall\epsilon > 0$ and $\forall R > 0$, $\exists\bar{m}$ such that $\forall m > \bar{m}$,

$$|\pi(x) - y_m| < R \quad \Rightarrow \quad \left|\frac{\pi(x)}{1+|\pi(x)|} - \frac{y_m}{1+|y_m|}\right| < \epsilon.$$

On the other hand, $\forall\epsilon > 0$, $\exists\bar{R} > 0$ such that $\forall R > \bar{R}$,

$$\int_{\mathbb{R}^N_+\setminus\Lambda_R(y_m)} |\tilde{\Psi}_{\bar{\sigma},y_m}(x)|^{2^*}dx = \int_{\mathbb{R}^N_+\setminus\Lambda_R(0)} |\tilde{\Psi}_{\bar{\sigma},0}(x)|^{2^*}dx < \epsilon. \tag{2.12}$$

Then chose $\epsilon > 0$ arbitrarily and fix $R > 0$ so that (2.12) is verified; if m is large enough, we get

$$\begin{aligned}\left|\beta(u_m) - \frac{y_m}{1+|y_m|}\right| &\le \int_{\mathbb{R}^N_+}\left|\frac{\pi(x)}{1+|\pi(x)|} - \frac{y_m}{1+|y_m|}\right| |u_m(x)|^{2^*}dx \\ &\le \int_{\mathbb{R}^N_+\setminus\Lambda_R(y_m)}\left|\frac{\pi(x)}{1+|\pi(x)|} - \frac{y_m}{1+|y_m|}\right| |\tilde{\Psi}_{\bar{\sigma},y_m}(x)|^{2^*}dx \\ &\quad + \int_{\Lambda_R(y_m)}\left|\frac{\pi(x)}{1+|\pi(x)|} - \frac{y_m}{1+|y_m|}\right| |\tilde{\Psi}_{\bar{\sigma},y_m}(x)|^{2^*}dx + o(1) \le 3\epsilon + o(1).\end{aligned}$$

Thus $|\beta(u_m)| \to 1$ as $m \to +\infty$, which again contradicts (2.10). Thus (2.11)(b) is also true, as desired. □

PROPOSITION 2.5 *Let $k > 0$, $k \in \mathbb{R}$. Then*

$$\inf\left\{\int_{\mathbb{R}^N_+}[|\nabla u|^2 + ku^2]dx : u \in W^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1,\right.$$
$$\left.\beta(u) = 0,\ \gamma(u) \ge \frac{1}{3}\right\} > \Sigma. \tag{2.13}$$

Proof. Clearly,

$$\inf\left\{\int_{\mathbb{R}^N_+}[|\nabla u|^2 + ku^2]dx : u \in W^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1,\ \beta(u) = 0,\ \gamma(u) \ge \frac{1}{3}\right\} \ge \Sigma.$$

Then to prove the proposition, it is sufficient to show that in the above relation the equality cannot be true. We argue by contradiction, so we suppose that there exists a sequence $\{u_n\}$, $u_n \in W^{1,2}(\mathbb{R}^N_+)$, such that

$$\begin{cases} \text{(a)} & |u_n|_{L^{2^*}(\mathbb{R}^N_+)} = 1, \qquad \beta(u_n) = 0, \qquad \gamma(u_n) \ge \frac{1}{3}, \\ \text{(b)} & \lim\limits_{n\to+\infty}\int_{\mathbb{R}^N_+}[|\nabla u_n|^2 + ku_n^2]dx = \Sigma. \end{cases} \tag{2.14}$$

Since $k > 0$, from

$$\Sigma = \lim_{n\to+\infty}\int_{\mathbb{R}^N_+}[|\nabla u_n|^2 + ku_n^2]dx \ge \lim_{n\to+\infty}\int_{\mathbb{R}^N_+}|\nabla u_n|^2dx \ge \Sigma,$$

we deduce $\lim_{n\to+\infty}\int_{\mathbb{R}^N_+}|\nabla u_n|^2dx=\Sigma$.

Hence by the uniqueness of the family of functions $\tilde{\Psi}_{\sigma,y}$ (defined in (2.4)) that realize Σ, we deduce that

$$u_n(x)=\tilde{\Psi}_{\sigma_n,y_n}(x)+w_n(x)\quad\forall x\in\mathbb{R}^N_+,$$

where $\sigma_n\in\mathbb{R}^+\backslash\{0\}$, $y_n\in\partial\mathbb{R}^N_+$, and $\{w_n\}$ is a sequence that goes strongly to zero in $W^{1,2}(\mathbb{R}^N_+)$ and $L^{2^*}(\mathbb{R}^N_+)$.

As in the proof of Proposition 2.4, we need to show that (up to subsequences)

$$\text{(2.15)}\qquad \text{(a)}\quad \lim_{n\to+\infty}\sigma_n=\bar{\sigma}>0,\ \bar{\sigma}\in\mathbb{R},\qquad \text{(b)}\quad \lim_{n\to+\infty}y_n=\bar{y}\in\partial\mathbb{R}^N_+.$$

In fact, if the relations in (2.15) hold, the proof can be concluded easily by observing that $\tilde{\Psi}_{\sigma_n,y_n}\to\tilde{\Psi}_{\bar{\sigma},\bar{y}}$ strongly in $W^{1,2}(\mathbb{R}^N_+)$ and $L^{2^*}(\mathbb{R}^N_+)$, which together with (2.14)(b) allows us to obtain the following impossible relation:

$$\begin{aligned}\Sigma&=\lim_{n\to+\infty}\int_{\mathbb{R}^N_+}[|\nabla u_n|^2+ku_n^2]dx=\lim_{n\to+\infty}\int_{\mathbb{R}^N_+}[|\nabla\tilde{\Psi}_{\sigma_n,y_n}|^2+k\tilde{\Psi}^2_{\sigma_n,y_n}]dx\\&=\int_{\mathbb{R}^N_+}[|\nabla\tilde{\Psi}_{\bar{\sigma},\bar{y}}(x)|^2+k\tilde{\Psi}^2_{\bar{\sigma},\bar{y}}(x)]dx>\int_{\mathbb{R}^N_+}|\nabla\tilde{\Psi}_{\bar{\sigma},\bar{y}}(x)|^2dx=\Sigma.\end{aligned}$$

To verify (2.15)(a), we first observe that σ_n must be bounded. In fact, if there were a subsequence (still denoted by σ_n) for which $\lim_{n\to+\infty}\sigma_n=+\infty$, we would obtain

$$\begin{aligned}\Sigma&=\lim_{n\to+\infty}\int_{\mathbb{R}^N_+}[|\nabla u_n|^2+ku_n^2]dx\geq\lim_{n\to+\infty}\left[\int_{\mathbb{R}^N_+}|\nabla\tilde{\Psi}_{\sigma_n,y_n}|^2+k\int_{B_{\sqrt{\sigma_n}(y_n)}}\tilde{\Psi}^2_{\sigma_n,y_n}(x)dx\right]\\&=\lim_{n\to+\infty}\left[\int_{\mathbb{R}^N_+}|\nabla\tilde{\Psi}_{1,0}(x)|^2dx+k\sigma_n\int_{B_1(0)}\tilde{\Psi}^2_{1,0}(x)dx\right]=+\infty.\end{aligned}$$

Then up to a subsequence, $\lim_{n\to+\infty}\sigma_n=\bar{\sigma}\in\mathbb{R}^+$ and $\bar{\sigma}>0$ can be deduced by arguing as in the proof of Proposition 2.4. Analogously, relation (2.15)(b) can be verified by following the argument used in the proof of Proposition 2.4 to prove (2.11)(b). □

3. Proof of the results. In what follows, we suppose that $a(x)$ verifies (1.1), and we use the following notations:

$$c_a=\inf\left\{\int_{\mathbb{R}^N_+}[|\nabla u|^2+a(x)u^2]dx : u\in V,\quad\beta(u)=0,\quad\gamma(u)=\frac{1}{3}\right\}$$

$$c_\infty=\inf\left\{\int_{\mathbb{R}^N_+}[|\nabla u|^2+a_\infty u^2]dx : u\in V,\quad\beta(u)=0,\quad\gamma(u)\geq\frac{1}{3}\right\}$$

$$\begin{aligned}c_{a-a_\infty}=\inf\bigg\{&\int_{\mathbb{R}^N_+}[|\nabla u|^2+(a(x)-a_\infty)u^2]dx : u\in\mathcal{D}^{1,2}(\mathbb{R}^N_+),\\&|u|_{L^{2^*}(\mathbb{R}^N_+)}=1,\quad\beta(u)=0,\quad\gamma(u)=\frac{1}{3}\bigg\}.\end{aligned}$$

By virtue of (1.1) and Propositions 2.4 and 2.5, we have

$$\text{(3.1)}\qquad \begin{cases}\text{(i)}\quad a_\infty=0\ \Rightarrow\ c_{a-a_\infty}=c_a>\Sigma,\quad c_\infty=\Sigma,\\ \text{(ii)}\quad a_\infty>0\ \Rightarrow\ c_a\geq c_{a-a_\infty}>\Sigma,\quad c_\infty>\Sigma.\end{cases}$$

We set

$$\bar{c} = \min\left\{\frac{c_{a-a_\infty} + \Sigma}{2}, \ \frac{\Sigma + S}{2}\right\} \tag{3.2}$$

and remark that

$$\Sigma < \bar{c} < S.$$

We denote by $\varphi(x)$ a function that belongs to $W_0^{1,2}(B_1(0))$ and has the following properties:

$$\begin{cases} \varphi \in \mathcal{C}_0^\infty(B_1(0)), \quad \varphi(x) > 0 \quad \forall x \in B_1(0), \\ \varphi \text{ is radially symmetric} \quad \text{and} \quad |x_1| < |x_2| \Rightarrow \varphi(x_1) > \varphi(x_2), \\ |\varphi|_{L^{2^*}(\mathbb{R}^N_+ \cap B_1(0))} = 1, \\ \Sigma < \int_{\mathbb{R}^N_+ \cap B_1(0)} |\nabla\varphi|^2 dx \equiv \bar{\Sigma} < \bar{c}. \end{cases} \tag{3.3}$$

The existence of such a φ follows from the properties of Σ and from (3.2). Moreover, if $a(x)$ satisfies (1.2), φ is supposed to have been chosen in such a way that the condition

$$\bar{\Sigma} < S - |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)} \tag{3.4}$$

is also fulfilled.

For every $\sigma > 0$ and $y \in \mathbb{R}^N$, we set

$$\varphi_{\sigma,y}(x) = \begin{cases} \sigma^{-(N-2)/4}\varphi\left(\frac{x-y}{\sqrt{\sigma}}\right), & x \in B_\sigma(y), \\ 0, & x \notin B_\sigma(y), \end{cases}$$

and we remark that

$$|\varphi_{\sigma,y}|_{L^{2^*}(\mathbb{R}^N)} = |\varphi_{\sigma,y}|_{L^{2^*}(B_\sigma(y))} = |\varphi|_{L^{2^*}(B_1(0))}.$$

Lemma 3.1. *Let $\alpha(x) \in L^{N/2}(\mathbb{R}^N_+)$ be a nonnegative function. Then the following relations hold:*

$$\begin{cases} \text{(a)} \quad \lim_{\sigma\to 0} \sup\left\{\int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx : y \in \partial\mathbb{R}^N_+\right\} = 0 \\ \text{(b)} \quad \lim_{\sigma\to +\infty} \sup\left\{\int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx : y \in \partial\mathbb{R}^N_+\right\} = 0 \\ \text{(c)} \quad \lim_{r\to +\infty} \sup\left\{\int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx : |y| = r, \quad \sigma > 0, \quad y \in \partial\mathbb{R}^N_+\right\} = 0. \end{cases} \tag{3.5}$$

Proof. Let $y \in \partial\mathbb{R}^N_+$ be chosen arbitrarily. Then $\forall\sigma > 0$,

$$\begin{aligned} \int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx &= \int_{\mathbb{R}^N_+ \cap B_\sigma(y)} \alpha(x)\varphi^2_{\sigma,y}(x)dx \\ &\le |\alpha|_{L^{N/2}(\mathbb{R}^N_+ \cap B_\sigma(y))}|\varphi_{\sigma,y}|^2_{L^{2^*}(\mathbb{R}^N_+ \cap B_\sigma(y))} = |\alpha|_{L^{N/2}(B_\sigma(y)\cap\mathbb{R}^N_+)}, \end{aligned}$$

so

$$\sup\left\{\int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx : y \in \partial\mathbb{R}^N_+\right\} \le \sup\left\{|\alpha|_{L^{N/2}(B_\sigma(y)\cap\mathbb{R}^N_+)} : y \in \partial\mathbb{R}^N_+\right\}. \tag{3.6}$$

On the other hand, $\forall y \in \mathbb{R}^N$, $\lim_{\sigma\to 0} |\alpha|_{L^{N/2}(\mathbb{R}^N_+ \cap B_\sigma(y))} = 0$, so (3.5)(a) follows from (3.6).

To prove (3.5)(b) let us fix $y \in \partial\mathbb{R}^N_+$ arbitrarily. Then $\forall \rho > 0$, $\forall \sigma > 0$,

$$\begin{aligned}\int_{\mathbb{R}^N_+} & \alpha(x)\varphi^2_{\sigma,y}(x)dx \\ &= \int_{\mathbb{R}^N_+\cap B_\rho(0)} \alpha(x)\varphi^2_{\sigma,y}dx + \int_{\mathbb{R}^N_+\setminus B_\rho(0)} \alpha(x)\varphi^2_{\sigma,y}(x)dx \\ &\leq |\alpha|_{L^{N/2}(B_\rho(0)\cap\mathbb{R}^N_+)}|\varphi_{\sigma,y}|^2_{L^{2^*}(B_\rho(0))} + |\alpha|_{L^{N/2}(\mathbb{R}^N_+\setminus B_\rho(0))}|\varphi_{\sigma,y}|^2_{L^{2^*}(\mathbb{R}^N_+\setminus B_\rho(0))} \\ &\leq |\alpha|_{L^{N/2}(\mathbb{R}^N_+\cap B_\rho(0))} \sup_{y\in\partial\mathbb{R}^N_+} |\varphi_{\sigma,y}|^2_{L^{2^*}(B_\rho(0))} + |\alpha|_{L^{N/2}(\mathbb{R}^N_+\setminus B_\rho(0))}.\end{aligned}$$

Moreover, $\forall y \in \mathbb{R}^N$, $\lim_{\sigma\to+\infty} |\varphi_{\sigma,y}|_{L^{2^*}(B_\rho(0))} = 0$, so we get

$$\lim_{\sigma\to+\infty} \sup\left\{\int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma,y}(x)dx : \ y \in \partial\mathbb{R}^N_+\right\} \leq |\alpha|_{L^{N/2}(\mathbb{R}^N_+\setminus B_\rho(0))},$$

and, letting $\rho \to \infty$, (3.5)(b) follows.

To verify (3.5)(c), we argue by contradiction, so we assume that there exist a sequence $\{y_n\}$, $y_n \in \partial\mathbb{R}^N_+$, and a sequence σ_n of positive numbers such that

$$\lim_{n\to+\infty} \int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma_n,y_n}(x)dx > 0 \quad \text{and} \quad |y_n| \to +\infty. \tag{3.7}$$

Because (3.5)(a)–(b) pass eventually to a subsequence, we can suppose that $\lim_{n\to+\infty} \sigma_n = \bar{\sigma}$. Then $|y_n| \to +\infty$ and $\alpha \in L^{N/2}(\mathbb{R}^N_+)$ imply $\lim_{n\to+\infty} |\alpha|_{L^{N/2}(\mathbb{R}^N_+\cap B_{\sigma_n}(y_n))} = 0$, from which we deduce

$$\lim_{n\to+\infty} \int_{\mathbb{R}^N_+} \alpha(x)\varphi^2_{\sigma_n,y_n}(x)dx \leq \lim_{n\to+\infty} \left[|\alpha|_{L^{N/2}(\mathbb{R}^N_+\cap B_{\sigma_n}(y_n))}|\varphi_{\sigma_n,y_n}|^2_{L^{2^*}(B_{\sigma_n}(y_n))}\right] = 0,$$

which contradicts (3.7). □

Lemma 3.2. *The following relations hold:*

$$\begin{cases} \text{(a)} \quad \lim_{\sigma\to 0} \sup\{\gamma(\varphi_{\sigma,y}) : y \in \partial\mathbb{R}^N_+\} = 0, \\ \text{(b)} \quad \lim_{\sigma+\infty} \inf\{\gamma(\varphi_{\sigma,y}) : y \in \partial\mathbb{R}^N_+, |y| \leq r\} = 1 \quad \forall r > 0, \\ \text{(c)} \quad (\beta(\varphi_{\sigma,y})|y)_{\mathbb{R}^N} > 0, \quad \forall y \in \partial\mathbb{R}^N_+\setminus\{0\}, \quad \forall \sigma > 0. \end{cases} \tag{3.8}$$

Proof. Let $y \in \partial\mathbb{R}^N_+$ be chosen arbitrarily. For any $\sigma > 0$, we have

$$\begin{aligned}
0 \le \gamma(\varphi_{\sigma,y}) &= \int_{\mathbb{R}^N_+} \left| \frac{\pi(x)}{1+|\pi(x)|} - \beta(\varphi_{\sigma,y}) \right| |\varphi_{\sigma,y}(x)|^{2^*} dx \\
&\le \int_{\mathbb{R}^N_+ \cap B_\sigma(y)} \left| \frac{\pi(x)}{1+|\pi(x)|} - \frac{y}{1+|y|} \right| |\varphi_{\sigma,y}(x)|^{2^*} dx + \left| \frac{y}{1+|y|} - \beta(\varphi_{\sigma,y}) \right| \\
&\le \int_{\mathbb{R}^N_+ \cap B_\sigma(y)} |\pi(x) - y| \, |\varphi_{\sigma,y}(x)|^{2^*} dx \\
&\quad + \left| \int_{\mathbb{R}^N_+} \left(\frac{y}{1+|y|} - \frac{\pi(x)}{1+|\pi(x)|} \right) |\varphi_{\sigma,y}(x)|^{2^*} dx \right| \\
&\le \sigma + \int_{\mathbb{R}^N_+} \left| \frac{y}{1+|y|} - \frac{\pi(x)}{1+|\pi(x)|} \right| |\varphi_{\sigma,y}(x)|^{2^*} dx \le 2\sigma.
\end{aligned}$$

Hence $0 \le \sup\{\gamma(\varphi_{\sigma,y}) : y \in \partial\mathbb{R}^N_+\} \le 2\sigma$, which letting $\sigma \to 0$, yields (3.8)(a).

In order to prove (3.8)(b), let us first show that $\forall y \in \partial\mathbb{R}^N_+$,

$$\lim_{\sigma \to +\infty} |\beta(\varphi_{\sigma,y})| = 0. \tag{3.9}$$

Since $\beta(\varphi_{\sigma,0}) = 0$ because of symmetry, we have

$$\begin{aligned}
|\beta(\varphi_{\sigma,y})| &= |\beta(\varphi_{\sigma,y}) - \beta(\varphi_{\sigma,0})| \\
&= \left| \int_{\mathbb{R}^N_+} \frac{\pi(x)}{1+|\pi(x)|} (\varphi^{2^*}_{\sigma,y}(x) - \varphi^{2^*}_{\sigma,0}(x)) dx \right| \\
&\le \int_{\mathbb{R}^N_+} \frac{|\pi(x)|}{1+|\pi(x)|} |\varphi^{2^*}_{\sigma,y}(x) - \varphi^{2^*}_{\sigma,0}(x)| dx \le \int_{\mathbb{R}^N_+} |\varphi^{2^*}_{1,\frac{y}{\sigma}}(x) - \varphi^{2^*}_{1,0}(x)| dx = O\left(\frac{1}{\sigma}\right),
\end{aligned}$$

which gives (3.9). Now fix $r > 0$ arbitrarily and let $y \in \partial\mathbb{R}^N_+$ so that $|y| \le r$. For any $\sigma > 0$, we have

$$\begin{aligned}
\gamma(\varphi_{\sigma,y}) &= \int_{\mathbb{R}^N_+} \left| \frac{\pi(x)}{1+|\pi(x)|} - \beta(\varphi_{\sigma,y}) \right| |\varphi_{\sigma,y}(x)|^{2^*} dx \\
&\le \int_{\mathbb{R}^N_+} \frac{\pi(x)}{1+|\pi(x)|} |\varphi_{\sigma,y}(x)|^{2^*} dx + |\beta(\varphi_{\sigma,y})| \le 1 + |\beta(\varphi_{\sigma,y})|,
\end{aligned}$$

from which, using (3.9), we deduce

$$\limsup_{\sigma \to +\infty} \inf\{\gamma(\varphi_{\sigma,y}) : y \in \partial\mathbb{R}^N_+, \ |y| \le r\} \le 1. \tag{3.10}$$

Now if

$$\liminf_{\sigma \to +\infty} \inf\{\gamma(\varphi_{\sigma,y}) : y \in \partial\mathbb{R}^N_+, \ |y| \le r\} < 1$$

holds, there exist a sequence of positive numbers $\{\sigma_n\}$ and a sequence of points $\{y_n\}$, $y_n \in \partial\mathbb{R}^N_+$, such that

$$\lim_{n \to +\infty} \gamma(\varphi_{\sigma_n,y_n}) < 1, \quad \sigma_n \xrightarrow[n \to +\infty]{} +\infty, \quad |y_n| \le r. \tag{3.11}$$

On the other hand, considering (3.9), we deduce $\forall\rho > 0$ that

$$\begin{aligned}\gamma(\varphi_{\sigma_n,y_n}) &= \int_{\mathbb{R}^N_+} \left|\frac{\pi(x)}{1+|\pi(x)|} - \beta(\varphi_{\sigma_n,y_n})\right| |\varphi_{\sigma_n,y_n}(x)|^{2^*} dx \\ &\geq \int_{\mathbb{R}^N_+} \frac{|\pi(x)|}{1+|\pi(x)|} |\varphi_{\sigma_n,y_n}(x)|^{2^*} dx - |\beta(\varphi_{\sigma_n,y_n})| \\ &\geq \int_{\mathbb{R}^N_+\setminus\Lambda_\rho(0)} \frac{|\pi(x)|}{1+|\pi(x)|} |\varphi_{\sigma_n,y_n}(x)|^{2^*} dx - o(1) \\ &\geq \frac{\rho}{(1+\rho)} \int_{\mathbb{R}^N_+\setminus\Lambda_\rho(0)} |\varphi_{\sigma_n,y_n}(x)|^{2^*} dx - o(1) \\ &\geq \frac{\rho}{(1+\rho)} \int_{\mathbb{R}^N_+\setminus\Lambda_{\frac{\rho}{\sigma_n}}(0)} |\varphi_{\sigma_n,y_n}(x)|^{2^*} dx - o(1),\end{aligned}$$

which implies that $\lim_{n\to+\infty}\gamma(\varphi_{\sigma_n,y_n}) \geq \rho/(1+\rho)$ $\forall\rho > 0$, and then

$$\lim_{n\to+\infty}\gamma(\varphi_{\sigma_n,y_n}) \geq 1,$$

which contradicts (3.11). Thus in (3.10), equality must hold, and since the above argument does not depend on the choice of r, (3.8)(b) is proven.

Finally, let us remark that if $0 \notin B_\sigma(y)$, (3.8)(c) is immediate. On the other hand, if $0 \in B_\sigma(y)$, $\forall x \in B_\sigma(y) \cap \mathbb{R}^N_+$ such that $(\pi(x)|y) < 0$, the point $\bar{x}$, symmetrical to $-x$ with respect to $\partial\mathbb{R}^N_+$, belongs to $B_\sigma(y)\cap\mathbb{R}^N_+$, $(\pi(\bar{x})|y) > 0$, and $\varphi_{\sigma,y}(\bar{x}) > \varphi_{\sigma,y}(x)$. Thus $\forall\sigma > 0$,

$$(\beta(\varphi_{\sigma,y})|y) = \int_{\mathbb{R}^N_+\cap B_\sigma(y)} \frac{(\pi(x)|y)}{1+|\pi(x)|} |\varphi_{\sigma,y}(x)|^{2^*} dx > 0,$$

as desired. □

Corollary 3.3. *Let $a(x)$ satisfy (1.1). Let $\epsilon > 0$ be a real number such that $\bar{\Sigma} + \epsilon < \bar{c}$. Then there exist real numbers $r > 0$ and $\sigma_1, \sigma_2 : 0 < \sigma_1 < 1/3 < \sigma_2$ such that*

$$\gamma(\varphi_{\sigma_1,y}) < \frac{1}{3}, \qquad \gamma(\varphi_{\sigma_2,y}) > \frac{1}{3} \quad \forall y \in \partial\mathbb{R}^N_+, \tag{3.12}$$

and

$$\sup\left\{\int_{\mathbb{R}^N_+}[|\nabla\varphi_{\sigma,y}(x)|^2 + (a(x)-a_\infty)\varphi^2_{\sigma,y}(x)]dx \ : \ (y,\sigma)\in\partial\mathcal{K}\right\} < \bar{\Sigma} + \frac{\epsilon}{2}, \tag{3.13}$$

where

$$\mathcal{K} = \{(y,\sigma) \in \partial\mathbb{R}^N_+ \times \mathbb{R}^+ : |y| \leq r, \ \sigma \in [\sigma_1,\sigma_2]\}. \tag{3.14}$$

Proof. By (3.8)(a) and (3.5)(a) with (1.1), there exists $\sigma_1 \in (0, 1/3)$ such that $\gamma(\varphi_{\sigma_1,y}) < 1/3$ $\forall y \in \partial\mathbb{R}^N_+$ and the relation

$$\int_{\mathbb{R}^N_+}[|\nabla\varphi_{\sigma,y}(x)|^2 + (a(x)-a_\infty)\varphi^2_{\sigma,y}(x)]dx < \bar{\Sigma} + \frac{\epsilon}{2} \tag{3.15}$$

holds when $\sigma = \sigma_1$ for any $y \in \partial\mathbb{R}^N_+$. Furthermore, (3.5)(c) with (1.1) allows us to choose $r > 0$ such that if $|y| = r$ and $y \in \partial\mathbb{R}^N_+$, (3.15) is satisfied for all $\sigma > 0$. Lastly, fixing r as chosen before, it is possible by (3.8)(b) and (3.5)(b) with (1.1) to find $\sigma_2 > 1/3$ for which $\gamma(\varphi_{\sigma_2,y}) > 1/3$ for any $y \in \partial\mathbb{R}^N_+$, $|y| \leq r$ and relation (3.15) holds with $\sigma = \sigma_2$ for any $y \in \partial\mathbb{R}^N_+$. Clearly, the set $\mathcal{K} = \{y \in \partial\mathbb{R}^N_+ : |y| \leq r\} \times [\sigma_1, \sigma_2]$ with r, σ_1, and σ_2 characterized as before is the desired set that satisfies (3.13). □

COROLLARY 3.4. *Let $a(x)$ satisfy (1.1) and $a_\infty > 0$. Let $\epsilon > 0$ be a real number chosen so that $\bar{\Sigma} + \epsilon < \bar{c}$ and, if (1.2) holds, $\bar{\Sigma} + \epsilon < S - |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)}$. Let σ_1, σ_2, and r be the numbers (that depend on ϵ) whose existence is stated in Corollary 3.3, and let $\mathcal{K}$ denote the set defined in (3.14). Then there exists a real number $\mathcal{A} > 0$ such that if $a_\infty \in (0, \mathcal{A})$, the relation*

$$\sup\left\{\int_{\mathbb{R}^N_+} [|\nabla\varphi_{\sigma,y}(x)|^2 + a(x)\varphi^2_{\sigma,y}(x)]dx : \quad (y,\sigma) \in \partial\mathcal{K}\right\} < \bar{\Sigma} + \epsilon < \bar{c} \tag{3.16}$$

holds, and if (1.2) is true,

$$\sup\left\{\int_{\mathbb{R}^N_+} [|\nabla\varphi_{\sigma,y}(x)|^2 + a(x)\varphi^2_{\sigma,y}(x)]dx : \quad (y,\sigma) \in \mathcal{K}\right\} < S \tag{3.17}$$

is also verified.

Proof. Since $\forall k > 0, \ \forall y \in \partial\mathbb{R}^N_+$,

$$\int_{\mathbb{R}^N_+} k\varphi^2_{\sigma,y}(x)dx = k\sigma^2|\varphi_{1,0}|^2_{L^2(B_1(0))},$$

relation (3.16) follows straightly from (3.13) when $a_\infty \in (0, \mathcal{A})$ with $\mathcal{A} = \epsilon(2\sigma_2^2|\varphi_{1,0}|^2_{L^2(B_1(0))})^{-1}$. Analogously, since $\forall y \in \partial\mathbb{R}^N_+$and $\forall\sigma > 0$

$$\int_{\mathbb{R}^N_+} (a(x) - a_\infty)\varphi^2_{\sigma,y}(x)dx \leq |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)}|\varphi_{\sigma,y}|^2_{L^{2^*}(\mathbb{R}^N_+)} = |a(x) - a_\infty|_{L^{N/2}(\mathbb{R}^N_+)},$$

when (1.2) is satisfied, (3.17) is true if $a_\infty \in (0, \mathcal{A})$. □

LEMMA 3.5. *Let $\mathcal{K}$ be the set defined in (3.14) with σ_1, σ_2, and r chosen as in Corollary 3.3. Then*

$$\exists(\bar{y}, \bar{\sigma}) \in \partial\mathcal{K} : \beta(\varphi_{\bar{\sigma},\bar{y}}) = 0, \quad \gamma(\varphi_{\bar{\sigma},\bar{y}}) \geq \frac{1}{3},$$
$$\exists(\hat{y}, \hat{\sigma}) \in \overset{\circ}{\mathcal{K}} \ : \beta(\varphi_{\hat{\sigma},\hat{y}}) = 0, \quad \gamma(\varphi_{\hat{\sigma},\hat{y}}) = \frac{1}{3}.$$

Proof. To prove the lemma, define the map $\vartheta : \partial\mathcal{K} \to \mathbb{R}^{N-1} \times \mathbb{R}$ by

$$\vartheta(y,\sigma) = (\beta(\varphi_{\sigma,y}), \gamma(\varphi_{\sigma,y}));$$

it is sufficient to show that its restriction to $\partial\mathcal{K}$ is homotopically equivalent to the identity map in $\mathbb{R}^{N-1} \times \mathbb{R}\backslash\{(0, 1/3)\}$.

Therefore, let us consider the homotopy $\Theta : [0,1] \times \partial\mathcal{K} \to \mathbb{R}^{N-1} \times \mathbb{R}$,

$$\Theta(t, y, \sigma) = (1-t)(y,\sigma) + t(\beta(\varphi_{\sigma,y}), \ \gamma(\varphi_{\sigma,y})). \tag{3.18}$$

Clearly Θ is continuous, $\Theta(0, y, \sigma) = (y, \sigma)$, $\Theta(1, y, \sigma) = \vartheta(y, \sigma)$, so it remains to show that

$$(0, 1/3) \notin \Theta(t, \partial\mathcal{K}) \quad \forall t \in [0, 1] \tag{3.19}$$

or, equivalently,

$$\Theta(t, y, \sigma) \neq (0, 1/3) \quad \forall (y, \sigma) \in \partial\mathcal{K} \quad \forall t \in [0, 1].$$

In fact, set $\partial\mathcal{K} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \mathcal{F}_3$ with $\mathcal{F}_1 = \{(y, \sigma) : |y| \leq r,\ \sigma = \sigma_1\}$, $\mathcal{F}_2 = \{(y, \sigma) : |y| \leq r,\ \sigma = \sigma_2\}$, and $\mathcal{F}_3 = \{(y, \sigma) : |y| = r,\ \sigma \in [\sigma_1, \sigma_2]\}$. If $(y, \sigma) \in \mathcal{F}_1$, by (3.12), we have $\forall t \in [0, 1]$

$$(1 - t)\sigma_1 + t\gamma(\varphi_{\sigma_1,y}) < (1 - t)\frac{1}{3} + t\frac{1}{3} = \frac{1}{3}.$$

Analogously, if $(y, \sigma) \in \mathcal{F}_2$, by (3.12), $\forall t \in [0, 1]$,

$$(1 - t)\sigma_2 + t\gamma(\varphi_{\sigma_2,y}) > (1 - t)\frac{1}{3} + t\frac{1}{3} = \frac{1}{3}.$$

If $(y, \sigma) \in \mathcal{F}_3$, then $|y| = r$ and $0 < \sigma_1 \leq \sigma \leq \sigma_2$, so using (3.8)(c), we obtain $\forall t \in [0, 1]$

$$((1 - t)y + t\beta(\varphi_{\sigma,y})|y) = (1 - t)|y|^2 + t(\beta(\varphi_{\sigma,y})|y) > 0. \qquad \square$$

Proof of Theorem 1.1. Let $\mathcal{A}$ be the positive number whose existence is stated in Corollary 3.4 and assume that $a_\infty \in (0, \mathcal{A})$. Let $\epsilon > 0$ be chosen as in Corollary 3.4. We recall that in this case the space H in which the functional E is defined is $W^{1,2}(\mathbb{R}^N_+)$, so the constraint is $V = \{u \in W^{1,2}(\mathbb{R}^N_+) : |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1\}$. In the following, we use the notation

$$E^c = \{u \in V : E(u) < c\}, \quad c \in \mathbb{R}.$$

We set

$$\begin{aligned} c_1 &= \inf\left\{E(u) : u \in V,\ \beta(u) = 0,\ \gamma(u) \geq \frac{1}{3}\right\}, \\ b_1 &= \sup\{E(\varphi_{\sigma,y}) : (y, \sigma) \in \partial\mathcal{K}\}, \\ b_2 &= \sup\{E(\varphi_{\sigma,y}) : (y, \sigma) \in \mathcal{K}\}, \end{aligned}$$

and we recall that at the beginning of the section, we defined the numbers c_a, c_∞, and c_{a-a_∞}. By (3.1)(ii), we have $\Sigma < c_\infty \leq c_1$, Lemma 3.5 gives $c_1 \leq E(\varphi_{\bar{\sigma},\bar{y}}) \leq b_1$, and $b_1 < \bar{\Sigma} + \epsilon < \bar{c} < S$ follows from (3.16) and the definition in (3.2). Furthermore, by (3.1)(ii) and the definition in (3.2) $\bar{c} < c_a$, Lemma 3.5 gives $c_a \leq E(\varphi_{\hat{\sigma},\hat{y}}) \leq b_2$, and if (1.2) holds, $b_2 < S$ by (3.17). To summarize,

$$\Sigma < c_1 \leq b_1 < \bar{\Sigma} + \epsilon < \bar{c} < S, \tag{3.20}$$

$$\bar{c} < c_a \leq b_2 \quad \text{and} \quad b_2 < S \quad \text{if (1.2) holds.} \tag{3.21}$$

First, we prove that the functional E constrained on V has a critical level in the energy range $(\Sigma, \bar{\Sigma} + \epsilon)$. Let us choose $\bar{\epsilon} > 0$ so that $\Sigma < c_1 - \bar{\epsilon} < b_1 + \bar{\epsilon} < \bar{\Sigma} + \epsilon$ and, arguing by contradiction, suppose that

$$\{u \in V : c_1 - \bar{\epsilon} \leq E(u) \leq b_1 + \bar{\epsilon}\ ,\ \nabla E_{|V}(u) = 0\} = \emptyset. \tag{3.22}$$

By Proposition 2.3 and (3.20), the pair (E, V) satisfies the Palais–Smale condition in $[c_1 - \bar{\epsilon}, b_1 + \bar{\epsilon}]$. Therefore, using a well-known deformation lemma (see, for instance, [S]), we find a continuous map $\eta : [0,1] \times V \to V$, and a positive number $\epsilon_0 < \bar{\epsilon}$ so that

$$\begin{aligned}\eta(0,u) &= u \quad \forall u \in V,\\ \eta(t,u) &= u \quad \forall u \in E^{c_1-\epsilon_0} \cup (V \backslash E^{b_1+\epsilon_0}) \quad \forall t \in [0,1],\\ E \circ \eta(t,u) &\leq E(u) \quad \forall t \in [0,1],\\ \eta(1, E^{b_1+\epsilon_0}) &\subset E^{c_1-\epsilon_0}.\end{aligned}$$

We remark that in particular

$$(y,\sigma) \in \partial\mathcal{K} \;\Rightarrow\; E(\varphi_{\sigma,y}) < b_1 \;\Rightarrow\; E(\eta(1,\varphi_{\sigma,y})) < c_1 - \epsilon_0. \tag{3.23}$$

Then let us define $\forall t \in [0,1]$ and $\forall (y,\sigma) \in \partial\mathcal{K}$ the map

$$\Gamma(t,y,\sigma) = \begin{cases} \Theta(2t,y,\sigma) & \forall t \in [0,1/2],\\ (\beta \circ \eta(2t-1,\varphi_{\sigma,y}),\ \gamma \circ \eta(2t-1,\varphi_{\sigma,y})) & \forall t \in [1/2,1],\end{cases}$$

where Θ is the map defined in (3.18). Γ is continuous, and $(0,1/3) \neq \Gamma(t,y,\sigma)\ \forall (y,\sigma) \in \partial\mathcal{K}\ \forall t \in [0,1/2]$ as a consequence of (3.19). Moreover, if $(y,\sigma) \in \partial\mathcal{K}$, the inequalities

$$E(\eta(2t-1,\varphi_{\sigma,y})) \leq E(\varphi_{\sigma,y}) \leq b_1 < \bar{c} < c_a = \inf\left\{E(u) :\ u \in V,\ \beta(u) = 0,\ \gamma(u) = \frac{1}{3}\right\}$$

$\forall t \in [1/2,1]$ hold. Thus we also have $(0,1/3) \neq \Gamma(t,y,\sigma)\ \forall (y,\sigma) \in \partial\mathcal{K}\ \ \forall t \in [1/2,1]$. Hence $(\tilde{y},\tilde{\sigma}) \in \partial\mathcal{K}$ must exist such that

$$\beta \circ \eta(1,\varphi_{\tilde{\sigma},\tilde{y}}) = 0, \quad \gamma \circ \eta(1,\varphi_{\tilde{\sigma},\tilde{y}}) \geq \frac{1}{3},$$

and then

$$E(\eta(1,\varphi_{\tilde{\sigma},\tilde{y}})) \geq \inf\{E(u) : u \in V,\ \beta(u) = 0,\ \gamma(u) \geq 1/3\} = c_1 > c_1 - \epsilon_0,$$

which contradicts (3.23), so (3.22) must be false.

Therefore, the functional E constrained on V must have at least one critical point $v \in V$ such that $\Sigma < E(v) < \bar{\Sigma} + \epsilon$, and by Lemma 2.2, v does not change sign. Therefore, because of the symmetry of E, v can be supposed positive.

Let us now assume that (1.2) holds. We shall prove that the functional E constrained on V has another critical level in the interval $(\bar{c}, S)$. As before, we argue by contradiction and suppose that, choosing $\epsilon' > 0$ such that $\bar{c} < c_a - \epsilon' < b_2 + \epsilon' < S$,

$$\{u \in V : c_a - \epsilon' \leq E(u) \leq b_2 + \epsilon',\ (\nabla E_{|V})(u) = 0\} = \emptyset. \tag{3.24}$$

By (3.21) and Proposition 2.3, the pair (E, V) satisfies the Palais–Smale condition in $(c_a - \epsilon', b_2 + \epsilon')$. Thus it is again possible to find a continuous map $\tilde{\eta} : [0,1] \times V \to V$ and a positive number $\epsilon_1 < \epsilon'$ so that

$$\begin{aligned}\tilde{\eta}(0,u) &= u \quad \forall u \in V,\\ \tilde{\eta}(t,u) &= u \quad \forall u \in E^{c_a-\epsilon_1} \cup (V \backslash E^{b_2+\epsilon_1})\ \forall t \in [0,1],\\ E \circ \tilde{\eta}(t,u) &\leq E(u) \quad \forall t \in [0,1],\\ \tilde{\eta}(1, E^{b_2+\epsilon_1}) &\subset E^{c_a-\epsilon_1}.\end{aligned}$$

We remark that in particular

$$(3.25)\qquad (y,\sigma)\in\mathcal{K}\ \Rightarrow\ E(\varphi_{\sigma,y})\le b_2\ \Rightarrow\ E(\tilde\eta(1,\varphi_{\sigma,y}))\le c_a-\epsilon_1.$$

Now define $\forall t\in[0,1]$ and $\forall(y,\sigma)\in\mathcal{K}$ the map

$$\tilde\Gamma(t,y,\sigma)=\begin{cases}\Theta(2t,y,\sigma), & t\in[0,\frac12],\\ (\beta\circ\tilde\eta(2t-1,\varphi_{\sigma,y}),\ \gamma\circ\tilde\eta(2t-1,\varphi_{\sigma,y})), & t\in[\frac12,1],\end{cases}$$

where Θ is the map defined in (3.18).

$\tilde\Gamma$ is continuous and $(0,1/3)\ne\tilde\Gamma(t,y,\sigma)$ $\forall(y,\sigma)\in\partial\mathcal{K}$ and $\forall t\in[0,1/2]$ as a consequence of (3.19). Moreover, since

$$(y,\sigma)\in\partial\mathcal{K}\ \Rightarrow\ E(\varphi_{\sigma,y})\le\bar\Sigma+\epsilon<\bar c<c_a-\epsilon_1\ \Rightarrow\ \tilde\eta(2t-1,\varphi_{\sigma,y})=\varphi_{\sigma,y}\quad\forall t\in\left[\frac12,1\right],$$

we have $\forall t\in[1/2,1],\forall(y,\sigma)\in\partial\mathcal{K}$

$$\tilde\Gamma(t,y,\sigma)=\tilde\Gamma\left(\frac12,y,\sigma\right)=\Theta(1,y,\sigma).$$

Then $(0,1/3)\ne\tilde\Gamma(t,y,\sigma)$ $\forall(y,\sigma)\in\partial\mathcal{K}$ $\forall t\in[1/2,1]$.

Hence a $(y^*,\sigma^*)\in\mathcal{K}$ must exist such that

$$\beta\circ\tilde\eta(1,\varphi_{\sigma^*,y^*})=0,\qquad \gamma\circ\tilde\eta(1,\varphi_{\sigma^*,y^*})=\frac13,$$

and then

$$E(\tilde\eta(1,\varphi_{\sigma^*,y^*}))\ge\inf\left\{E(u):u\in V,\ \beta(u)=0,\ \gamma(u)=\frac13\right\}=c_a>c_a-\epsilon_1,$$

which contradicts (3.25), so (3.24) must be false. Therefore, the functional E constrained on V has at least one critical point $u\in V$ such that $\bar c<E(u)<S$. Clearly, $u\ne v$ because $E(v)<\bar c<E(u)$, and using Lemma 2.2, we deduce $u>0$, concluding the proof. □

Remark 3.6. If we fix a function $\alpha(x)\in L^{N/2}(\mathbb{R}^N_+)$, $|\alpha|_{L^{N/2}(\mathbb{R}^N_+)}\ne0$, and consider the family of functions $a_\lambda(x)=\lambda+\alpha(x)$, $\lambda>0$, Theorem 1.1 insures the existence of a positive solution v_λ of (P) if λ is suitably small. Furthermore, by construction, $\Sigma<E(v_\lambda/|v_\lambda|_{L^{2^*}(\mathbb{R}^N_+)})<\bar\Sigma+\epsilon$, and since ϵ can be taken arbitrarily small and $\bar\Sigma$ taken arbitrarily near to Σ, it is easy to derive $\lim_{\lambda\to0}E(v_\lambda/|v_\lambda|_{L^{2^*}(\mathbb{R}^N_+)})=\Sigma$.

If in addition $\alpha(x)$ verifies the relation $|\alpha|_{L^{N/2}(\mathbb{R}^N_+)}<(1-2^{-2/N})S$, then for λ small, (P) has at least another solution u_λ, and, denoted by

$$c_\alpha=\inf\left\{\int_{\mathbb{R}^N_+}[|\nabla u|^2+\alpha(x)u^2]dx:\ u\in\mathcal{D}^{1,2}(\mathbb{R}^N_+),\ |u|_{L^{2^*}(\mathbb{R}^N_+)}=1,\right.$$
$$\left.\beta(u)=0,\ \gamma(u)=\frac13\right\},$$

the relation

$$\liminf_{\lambda\to0}E\left(\frac{u_\lambda}{|u_\lambda|_{L^{2^*}(\mathbb{R}^N_+)}}\right)>\bar c=\frac{c_\alpha+\Sigma}{2}$$

holds. □

Proof of Theorem 1.2. Let $\epsilon > 0$ be chosen in such a way that $\bar{\Sigma} + \epsilon < \bar{c}$ and the claim of Corollary 3.3 is true. Let c_a be defined as in the beginning of the section, and let us recall that in this case the space H in which the functional E is defined is $\mathcal{D}^{1,2}(\mathbb{R}^N_+)$ and the constraint is $V = \{u \in \mathcal{D}^{1,2}(\mathbb{R}^N_+) : |u|_{L^{2^*}(\mathbb{R}^N_+)} = 1\}$. If we set

$$\hat{b}_2 = \sup\{E(\varphi_{\sigma,y}) : (y,\sigma) \in \mathcal{K}\},$$

we have by (1.2) that $\hat{b}_2 < S$. Therefore, using (3.1)(i), the definition in (3.2), and Lemma 3.5, we obtain

$$\Sigma < \bar{c} < c_a \leq E(\varphi_{\hat{\sigma},\hat{y}}) \leq \hat{b}_2 < S.$$

Then arguing exactly as in the second part of the proof of Theorem 1.1, it is possible to prove the existence of a critical point of E on V which corresponds to a critical level in the energy range $(\bar{c}, S)$, and then it is positive by Lemma 2.2. □

REFERENCES

[BC] V. Benci and G. Cerami, *Existence of positive solutions of the equation* $-\Delta u + a(x)u = u^{(N+2)/(N-2)}$ *in* $\mathbb{R}^N$, J. Funct. Anal., 88 (1990), pp. 90–117.

[BL] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations* I: *Existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–346.

[BN] H. Brezis and L. Nirenberg, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math., 36 (1983), pp. 437–477.

[P] D. Passaseo, *Esistenza e molteplicità di soluzioni positive per l'equazione* $-\Delta u + (a(x) + \lambda)u = u^{(N+2)/(N-2)}$ *in* $\mathbb{R}^N$, Pubbl. Dip. Mat. Univ. Pisa, 565 (1990) (in Italian).

[S] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, 1990.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 886–896, July 1997

ON THE INVERSE PROBLEM OF A FOURTH-ORDER SELF-ADJOINT BINOMIAL OPERATOR[*]

ALAN ELCRAT[†] AND VASSILIS G. PAPANICOLAOU[†]

Abstract. Let L be the binomial operator

$$L = \left(\frac{d}{dx}\right)^4 + q(x)$$

acting on $(0, \pi)$ with Dirichlet boundary conditions. We study the associated inverse spectral problem under the assumptions that q is symmetric, i.e., $q(\pi - x) = q(x)$.

Our analysis is inspired by the well-known work of Borg for the Sturm–Liouville case. We first derive the eigenfunction asymptotics by an approach that is different from the ones used in the second-order case (WKB, etc.). These asymptotics are then used to obtain the local uniqueness of the inverse problem.

Key words. fourth-order binomial operator, Dirichlet boundary conditions, Green function, eigenfunction asymptotics, inverse spectral problem

AMS subject classifications. 34B05, 34L20

PII. S0036141095290938

1. Introduction.

Consider the Dirichlet eigenvalue problem on $(0, \pi)$:

$$Lu = u'''' + q(x)u = \lambda u, \qquad u(0) = u''(0) = u(\pi) = u''(\pi) = 0 \tag{1}$$

(L is a so-called *binomial* operator), where $q \in C[0, \pi]$ is symmetric, namely, $q(\pi - x) = q(x)$, and without loss of generality,

$$\int_0^\pi q(x)dx = 0. \tag{2}$$

Let λ_n, $n = 1, 2, 3, \ldots$, be the eigenvalues of (1) (there may be finitely many n's such that $\lambda_{n+1} = \lambda_n$) and $\phi_n(x)$, $n = 1, 2, 3, \ldots$, be the corresponding normalized eigenfunctions. A classical inverse question here is whether q is uniquely determined from the spectrum $\{\lambda_n\}_{n=1}^\infty$.

For the corresponding second-order (i.e., the Sturm–Liouville) case, there are many ways to obtain the uniqueness result and, furthermore, reconstruct q. Borg [2] was the first to obtain a general uniqueness result. (He also reconstructed q by a successive-approximation procedure.) Soon after, Levinson [14] gave a very short and elegant uniqueness proof using contour integration and large-$|\lambda|$ asymptotics. (For a brief description of Levinson's idea, also see [10].) The reconstruction of q from a given spectral function was accomplished by Gelfand and Levitan [7] with the help of the celebrated (Volterra-type) transformation operator. (See [6] for the adaptation of the method of Gelfand and Levitan in the case where two spectra are given.) Krein followed a different approach, where the inverse problem is transformed to a moment problem. (For a nice exposition of Krein's method in English, see [13].) Another way to solve the inverse problem is to first obtain evolution equations for the eigenvalues

[*] Received by the editors August 30, 1995; accepted for publication (in revised form) February 22, 1996.
http://www.siam.org/journals/sima/28-4/29093.html

[†] Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033 (arelcrat@twsuvm.uc.twsu.edu, papanico@cs.twsu.edu).

(with respect to the interval) and then after solving these equations recover q via a trace formula (see [1]), although this method is more popular in the case where q is periodic and the problem is considered on the full line (see, for example, [4]). For a survey on analytic methods for solving inverse problems, see [16].

We believe that among the methods mentioned above, the most promising for (1) is Borg's original approach. It seems that none of the other methods mentioned work for (1) despite the facts that trace formulas are still valid (see, for example, [18]) and the Gelfand–Levitan approach works satisfactorily in the case of two unknown coefficients (see, for example, [9] and [15]). The recent studies of Zachary [23] for the higher-order inverse-scattering problem and Yurko [20], [21], [22], where certain higher-order *non-self-adjoint* binomial inverse problems are solved, suggest that we should expect uniqueness results.

To apply Borg's approach, we need (eigenvalue and) eigenfunction asymptotics for (1). The behavior of the large eigenvalues is known (see, for example, [18]). Our first task, then (in section 2 below), is to obtain the eigenfunction asymptotics. Notice that to derive these asymptotics, we had to introduce a new method since the standard approaches used in the second-order case (WKB, etc.) do not work in the fourth-order case because in the latter there are always solutions such as $\exp\left(\lambda^{1/4}\right)$ growing; hence it is difficult to isolate the eigenfunctions, which are bounded in λ. In section 3, we use the eigenfunction asymptotics to attack the inverse problem. At the end of the paper, we outline a reconstruction procedure based on a work of Hald (see [11]). We believe that the analysis presented in this paper extends to other equations, including higher-dimensional ones; therefore, there are various potential applications (e.g., in elasticity).

2. The eigenfunction asymptotics. For the eigenvalues of (1), it is established (see, for example, [18]) that if $q \in C^r[0,\pi]$, where $r = 0$, 1, 2 or 3, then

$$\lambda_n - n^4 = O\left(n^{-r}\right). \tag{3}$$

Notice that n^4, $n = 1, 2, 3, \ldots$, are the eigenvalues of the *unperturbed* case, namely, when $q \equiv 0$. In this case, the corresponding eigenfunctions are $\psi_n(x) = (2/\pi)^{1/2} \sin nx$.

Our method for computing the asymptotics of the eigenfunctions $\phi_n(x)$ of (1) as $n \to \infty$ is inspired by an approach that we found in some works of Karpeshina (see, for example, [12]).

Let $G(x, y; \lambda)$ be the Green function associated with (1). By definition, $G(x, y; \lambda)$ is the integral kernel of the operator $(L - \lambda)^{-1}$; therefore, we have the eigenfunction expansion

$$G(x, y; \lambda) = \sum_{j=1}^{\infty} \frac{\phi_j(x)\phi_j(y)}{\lambda_j - \lambda}. \tag{4}$$

Thus $G(x, y; \lambda)$ is meromorphic in λ with simple poles at the Dirichlet eigenvalues.

In the unperturbed case, the Green function becomes

$$K(x, y; \lambda) = \begin{cases} \frac{1}{2s^2}\left[\frac{\sin(sx)\sin s(\pi - y)}{s\sin(s\pi)} - \frac{\sinh(sx)\sinh s(\pi - y)}{s\sinh(s\pi)}\right] & \text{if } x \leq y, \\ \\ \frac{1}{2s^2}\left[\frac{\sin(sy)\sin s(\pi - x)}{s\sin(s\pi)} - \frac{\sinh(sy)\sinh s(\pi - x)}{s\sinh(s\pi)}\right] & \text{if } x \geq y, \end{cases} \tag{5}$$

where $s = \lambda^{1/4}$. Observe that $K(x, y; \lambda)$ is a meromorphic function of $s^4 = \lambda$. Hence no matter what fourth root we choose for s, the value of $K(x, y; \lambda)$ is the same.

It is easy to see that $G(x,y;\lambda)$ is the unique solution (as long as $\lambda \neq \lambda_n$) of the integral equation

$$G(x,y;\lambda) = K(x,y;\lambda) - \int_0^\pi K(x,\xi;\lambda)q(\xi)G(\xi,y;\lambda)d\xi. \tag{6}$$

By iterating (6), we obtain a formal (perturbation) series for $G(x,y;\lambda)$, namely,

$$G(x,y;\lambda) = \sum_{m=0}^{\infty}(-1)^m G_m(x,y;\lambda), \tag{7}$$

where

$$\begin{aligned} &G_0(x,y;\lambda) = K(x,y;\lambda), \\ &G_m(x,y;\lambda) = \textstyle\int_0^\pi K(x,\xi;\lambda)q(\xi)G_{m-1}(\xi,y;\lambda)d\xi, \quad m \geq 1. \end{aligned} \tag{8}$$

Notice that if $m \geq 1$, $G_m(x,y;\lambda)$ also can be written as

$$G_m(x,y;\lambda) = \int_0^\pi \cdots \int_0^\pi K(x,\xi_1;\lambda)q(\xi_1)K(\xi_1,\xi_2;\lambda)\cdots q(\xi_m)K(\xi_m,y;\lambda)d\xi_1\cdots d\xi_m. \tag{9}$$

We need to establish the (absolute and uniform) convergence of the series in (7) for certain values of λ. (Notice that if this series converges, then it obviously satisfies (6).) First, we introduce the complex variable s (as in the unperturbed case) so that

$$\lambda = s^4.$$

Next, we consider a family of contours in the s-plane as follows:

$$C_n = \{s = \sigma + i\tau : \, |\sigma - n| \leq 1/2 \text{ and } |\tau| = 1, \text{ or } |\sigma - n| = 1/2 \text{ and } |\tau| \leq 1\},$$

i.e., C_n is a rectangle of horizontal side 1 and vertical side 2, centered at n.

PROPOSITION 1. *If $s \in C_n$ and n is large enough so that*

$$\delta_n \overset{\text{def}}{=} \frac{150\pi \, \|q\|_\infty}{(n-1/2)^3} \leq \frac{1}{2}, \tag{10}$$

then the series in (7) *converges absolutely and uniformly on* $[0,\pi]\times[0,\pi]\times C_n$.

Proof. First, we notice that if $s \in C_n$, then

$$\left|\frac{\sin(sx)\sin s(\pi-y)}{\sin(s\pi)}\right| \leq \left(\frac{e^\pi+1}{2}\right)^2 < 146,$$

and for $x \leq y$ (reminder: σ is the real part of s),

$$\left|\frac{\sinh(sx)\sinh s(\pi-y)}{\sinh(s\pi)}\right| \leq \frac{e^{\sigma(x-y)}+3}{2\,(1-e^{-\sigma\pi})} < 4.$$

Therefore, (5) implies that

$$|K(x,y;\lambda)| < \frac{150}{|s|^3} \leq \frac{150}{(n-1/2)^3}.$$

Thus (9) gives

$$|G_m(x,y;\lambda)| \leq 150\pi\delta_n^m, \tag{11}$$

and since $\delta_n \leq 1/2$, the absolute and uniform convergence of the series in (7) follows immediately. □

We now introduce the contours Γ_n in the λ-plane that correspond to the C_n's, namely,

$$\Gamma_n = \left\{\lambda = s^4 : s \in C_n\right\}.$$

Notice that the length $l\left(\Gamma_n\right)$ of Γ_n satisfies

$$l\left(\Gamma_n\right) \leq cn^3, \tag{12}$$

where c is some (positive) constant.

If n is sufficiently large, a consequence of Proposition 1 above is that

$$\frac{1}{2\pi i}\oint_{\Gamma_n} G(x,y;\lambda)d\lambda = \sum_{m=0}^{\infty}(-1)^m\frac{1}{2\pi i}\oint_{\Gamma_n} G_m(x,y;\lambda)d\lambda.$$

Using the eigenfunction expansion (4) and the fact that (3) implies that Γ_n encloses exactly one λ_n and exactly one eigenvalue of the unperturbed problem, namely n^4, we get

$$\phi_n(x)\phi_n(y) = \frac{2}{\pi}\sin nx \sin ny - \sum_{m=1}^{\infty}(-1)^m\frac{1}{2\pi i}\oint_{\Gamma_n} G_m(x,y;\lambda)d\lambda. \tag{13}$$

Next, we give a bound for the sum in (13). Using (11) and (12), we obtain the estimate

$$\left|\sum_{m=2}^{\infty}(-1)^m\frac{1}{2\pi i}\oint_{\Gamma_n} G_m(x,y;\lambda)d\lambda\right| \leq 75cn^3\sum_{m=2}^{\infty}\delta_n^m = 75cn^3\frac{\delta_n^2}{1-\delta_n}.$$

Thus by (10), there is a constant M (depending on $\|q\|_\infty$) such that

$$\left|\sum_{m=2}^{\infty}(-1)^m\frac{1}{2\pi i}\oint_{\Gamma_n} G_m(x,y;\lambda)d\lambda\right| \leq \frac{M}{n^3}. \tag{14}$$

We now treat the first term of the series in (13). By (8) or (9), we obtain

$$\frac{1}{2\pi i}\oint_{\Gamma_n} G_1(x,y;\lambda)d\lambda = \frac{1}{2\pi i}\oint_{\Gamma_n}\int_0^\pi K(x,\xi;\lambda)q(\xi)K(\xi,y;\lambda)d\xi d\lambda.$$

The eigenfunction expansion of $K(x,y;\lambda)$ together with the residue theorem give

$$\frac{1}{2\pi i}\oint_{\Gamma_n} G_1(x,y;\lambda)d\lambda = -\frac{2}{\pi}\int_0^\pi \sin n\xi\left[K^n(\xi,y;n^4)\sin nx + K^n(x,\xi;n^4)\sin ny\right]q(\xi)d\xi,$$

where we have set

$$K^n(x,y,\lambda) = \frac{2}{\pi}\sum_{\substack{j=1\\ j\neq n}}^{\infty}\frac{\sin nx \sin ny}{j^4-\lambda}.$$

It follows that

$$\left|\frac{1}{2\pi i}\oint_{\Gamma_n} G_1(x,y;\lambda)d\lambda\right| \le \frac{8}{\pi}\,\|q\|_\infty \sum_{\substack{j=1\\ j\neq n}}^{\infty} \frac{1}{|j^4-n^4|}. \tag{15}$$

Lemma 1. As $n\to\infty$,

$$\sum_{\substack{j=1\\ j\neq n}}^{\infty} \frac{1}{|j^4-n^4|} = O\left(\frac{\ln n}{n^3}\right).$$

Proof. We split the series into three terms:

$$\sum_{j=1}^{n-1}\frac{1}{n^4-j^4}+\sum_{j=n+1}^{2n}\frac{1}{j^4-n^4}+\sum_{j=2n+1}^{\infty}\frac{1}{j^4-n^4}. \tag{16}$$

The last term is $O(n^{-3})$. Regarding the first term, we have

$$\sum_{j=1}^{n-1}\frac{1}{n^4-j^4}=\frac{1}{n^3}\sum_{j=1}^{n-1}\frac{1}{1-(j/n)^4}\frac{1}{n}\le\frac{1}{n^3}\int_0^{1-(1/n)}\frac{dx}{1-x^4}+\frac{1}{n^4-(n-1)^4}.$$

Thus

$$\sum_{j=1}^{n-1}\frac{1}{n^4-j^4}=O\left(\frac{\ln n}{n^3}\right).$$

The second term in (16) can be treated similarly. □

Using (14), (15), and Lemma 1 in (13), we arrive at the important formula

$$\phi_n(x)\phi_n(y)=\frac{2}{\pi}\sin nx\sin ny+O\left(\frac{\ln n}{n^3}\right).$$

Let us make the meaning of this formula more precise. For a fixed positive number Q, there is a $d_1>0$ (depending only on Q) such that if $\|q\|_\infty\le Q$, then

$$\left|\phi_n(x)\phi_n(y)-\frac{2}{\pi}\sin nx\sin ny\right|\le\frac{d_1\ln n}{n^3}\,\|q\|_\infty\quad\text{for all } n\ge 2 \tag{17}$$

and d_1 remains bounded as $Q\searrow 0$.

We can go a little further. Consider the linear operator

$$(Tf)(x)=f(\pi-x).$$

The eigenspace V corresponding to an eigenvalue of (1) can have dimension 1 or 2, and T maps V into V. Since $T^2=I$, $T|_V$ (T restricted on V) is diagonalizable (in fact, it is symmetric), and its eigenvalues are $+1$ or -1. Thus T has a complete set of eigenvectors, and each eigenvector ϕ satisfies $\phi(\pi-x)=\pm\phi(x)$. It follows that for each eigenfunction ϕ_n of (1), we can always assume that

$$\phi_n(\pi-x)=c_n\phi_n(x)\quad\text{where } c_n=\pm 1. \tag{18}$$

Furthermore, by substituting $y = \pi - x$ in (17) and using (18), we get that if n is sufficiently large, then

$$\phi_n(\pi - x) = (-1)^{n+1}\phi_n(x). \tag{19}$$

If we set $y = x$ in (17), we obtain

$$\phi_n(x)^2 = \frac{2}{\pi}\sin^2 nx + O\left(\frac{\ln n}{n^3}\right).$$

On the other hand, if we freeze y, (17) implies that (up to an error of size $O(n^{-3}\ln n)$) $\phi_n(x)$ changes sign like $\sin nx$. We have therefore established the main result of this section.

THEOREM 1. *Let Q be a fixed positive number. Then there is a $d_2 > 0$ (depending only on Q) such that if $\|q\|_\infty \leq Q$, the eigenfunctions of* (1) *satisfy*

$$\left|\phi_n(x) - \sqrt{\frac{2}{\pi}}\sin nx\right| \leq \frac{d_2 \ln n}{n^3}\|q\|_\infty, \quad n \geq 2,$$

for all $x \in [0, \pi]$. Furthermore, d_2 remains bounded as $Q \searrow 0$.

3. The inverse problem. Theorem 1 enables us to attack the inverse problem associated with (1). We follow Borg's original approach (see [2]).

Let us consider the two eigenvalue problems

$$u'''' + q(x)u = \lambda u, \qquad u(0) = u''(0) = u(\pi) = u''(\pi) = 0 \tag{20}$$

and

$$u'''' + \widetilde{q}(x)u = \lambda u, \qquad u(0) = u''(0) = u(\pi) = u''(\pi) = 0, \tag{21}$$

where q and $\widetilde{q}$ are continuous and symmetric and without loss of generality,

$$\int_0^\pi q(x)dx = \int_0^\pi \widetilde{q}(x)dx = 0. \tag{22}$$

If a is some quantity associated with (20), the corresponding quantity for (21) is denoted by $\widetilde{a}$. Let λ_n and $\widetilde{\lambda}_n$, $n = 1, 2, 3, \ldots$, be the eigenvalues of (20) and (21), respectively. Here we assume that

$$\lambda_n = \widetilde{\lambda}_n \quad \text{for all } n.$$

Then if $\phi_n(x)$ and $\widetilde{\phi}_n(x)$, $n = 1, 2, 3, \ldots$, are the corresponding normalized eigenfunctions, we have

$$\phi_n'''' + q(x)\phi_n = \lambda_n\phi_n \quad \text{and} \quad \widetilde{\phi}_n'''' + \widetilde{q}(x)\widetilde{\phi}_n = \lambda_n\widetilde{\phi}_n.$$

If we multiply the first equation by $\widetilde{\phi}_n$ and the second by ϕ_n and then subtract the second from the first, we obtain

$$\phi_n''''\widetilde{\phi}_n - \widetilde{\phi}_n''''\phi_n = (\widetilde{q} - q)\,\phi_n\widetilde{\phi}_n.$$

Now we integrate from 0 to π and use integration by parts and the boundary conditions that ϕ_n and $\widetilde{\phi}_n$ satisfy. Thus we get

$$\int_0^{\pi} [\widetilde{q}(x) - q(x)]\, \phi_n(x)\widetilde{\phi}_n(x)dx = 0, \quad n = 1, 2, 3, \ldots. \tag{23}$$

Let $L_s^2(0, \pi)$ be the Hilbert space of all L^2 symmetric functions on $(0, \pi)$, namely,

$$L_s^2(0, \pi) = \left\{f \in L^2(0, \pi) : \, f(\pi - x) = f(x) \text{ a.e. } x\right\}.$$

Then by (23), the uniqueness part for the inverse problem is related to the following question: how much of $L_s^2(0, \pi)$ does the set $\{\phi_n\widetilde{\phi}_n\}_{n=1}^{\infty}$ span?

If in particular the set $\{\phi_n\widetilde{\phi}_n\}_{n=1}^{\infty}$ is complete in $L_s^2(0, \pi)$, then we have uniqueness since (23) then implies that we must have $\widetilde{q}(x) \equiv q(x)$.

Following Borg, we set

$$U_n(x) = \sqrt{\frac{2}{\pi}} \left(\int_0^{\pi} \phi_n\widetilde{\phi}_n\right) - (2\pi)^{1/2}\, \phi_n(x)\widetilde{\phi}_n(x) \tag{24}$$

(so that the average of $U_n(x)$ on $(0, \pi)$ is 0). Then because of (22), formula (23) is equivalent to

$$\int_0^{\pi} [\widetilde{q}(x) - q(x)]\, U_n(x)dx = 0, \quad n = 1, 2, 3, \ldots. \tag{25}$$

Theorem 1 implies that for fixed Q, there is a $d_3 > 0$ (depending only on Q) such that if $\|q\|_{\infty}, \|\widetilde{q}\|_{\infty} \leq Q$, then

$$\left|U_n(x) - \sqrt{\frac{2}{\pi}} \cos(2nx)\right| \leq \frac{d_3 Q\,(1 + \ln n)}{n^3} \quad \text{for all } n. \tag{26}$$

(Notice that $\{1/\sqrt{\pi}\} \cup \{(2/\pi)^{1/2} \cos 2nx\}_{n=1}^{\infty}$ is an orthonormal basis of $L_s^2(0, \pi)$.) By (18), it follows that

$$U_n(\pi - x) = c_n\widetilde{c}_n U_n(x) = \pm U_n(x).$$

If $n \geq n_0(Q)$, formula (19) implies that $c_n = \widetilde{c}_n = (-1)^{n+1}$. Therefore,

$$U_n(\pi - x) = U_n(x) \quad \text{for all } n \geq n_0.$$

Hence $U_n \in L_s^2(0, \pi)$ if $n \geq n_0$ (and $U_n \perp L_s^2(0, \pi)$ or $U_n \in L_s^2(0, \pi)$ if $n < n_0$).

Next, we consider the expansions

$$U_n(x) = \sum_{k=1}^{\infty} r_{nk} \sqrt{\frac{2}{\pi}} \cos 2kx, \quad n \geq n_0,$$

where, of course,

$$r_{nk} = \int_0^{\pi} U_n(x) \sqrt{\frac{2}{\pi}} \cos(2kx)\, dx.$$

We can also set $r_{nk} = 0$ if $n < n_0$. Introduce the operator R acting on $l_2(\mathbf{N})$ with matrix $(r_{nk})_{1 \leq k,n < \infty}$. By (26), R is a bounded operator on $l_2(\mathbf{N})$ (also see Lemma 2 below). We write

$$R = I + T, \tag{27}$$

where

$$T = (t_{nk})_{1 \leq k,n < \infty}\,, \quad t_{nk} = r_{nk} - \delta_{nk}.$$

(δ_{nk} is the Kronecker delta.)

Lemma 2. *The operator T defined above is compact on $l_2(\mathbf{N})$. Furthermore, if Q is sufficiently small* (*where, as usual, Q is an upper bound of $\|q\|_\infty$ and $\|\widetilde{q}\|_\infty$*), *then $\|T\| < 1$.*

Proof. (In fact, we prove a slightly stronger statement). We have

$$\sum_{n=1}^{\infty}\sum_{k=1}^{\infty} t_{nk}^2 = \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} (r_{nk} - \delta_{nk})^2 = \sum_{n=1}^{\infty}\left(1 - 2r_{nn} + \sum_{k=1}^{\infty} r_{nk}^2\right). \tag{28}$$

Since

$$r_{nn} = \int_0^\pi U_n(x)\sqrt{\frac{2}{\pi}}\cos(2nx)\,dx,$$

formula (26) implies that

$$|r_{nn} - 1| \leq \frac{d_4\,(1 + \ln n)}{n^3}\,\|q\|_\infty\,. \tag{29}$$

Next, we observe that the Parseval equation gives

$$\sum_{k=1}^{\infty} r_{nk}^2 = \int_0^\pi U_n^2(x)dx.$$

Thus again using (26), we obtain

$$\left|\sum_{k=1}^{\infty} r_{nk}^2 - 1\right| \leq \frac{d_5\,(1 + \ln n)}{n^3}\,\|q\|_\infty\,.$$

Using this and (29) in (28), we obtain that the double series converges. □

Notice that if $\|T\| < 1$, then (27) implies that R is invertible, and this means that the set $\{1\} \cup \{U_n(x)\}_{n=1}^\infty$ is complete in $L_s^2(0,\pi)$. Thus we have established our second main result, namely the following.

Theorem 2. *There is a $Q > 0$ such that if problems* (20) *and* (21) *have the same spectrum* (*where q and $\widetilde{q}$ are symmetric*), *and if $\|q\|_\infty\,, \|\widetilde{q}\|_\infty \leq Q$, then $q(x) \equiv \widetilde{q}(x)$.*

Remarks. (a) Notice that for our analysis, we needed only q, $\widetilde{q} \in L^\infty(0,\pi)$.

(b) We believe that our analysis can provide a precise numerical value for Q.

(c) If q and $\widetilde{q}$ are not symmetric, then we believe that a theorem analogous to Theorem 2 can be proved using this same method under the assumptions that the operators have two spectra in common and that the norms $\|q\|_\infty$ and $\|\widetilde{q}\|_\infty$ are sufficiently small. In the special case where q and $\widetilde{q}$ have the same Dirichlet spectrum on $(0,\pi)$—say σ_D—and the same spectrum—denoted by σ_{DN}—that corresponds to the boundary conditions $u(0) = u''(0) = u'(\pi) = u'''(\pi) = 0$, we can consider the symmetric extensions q_s and $\widetilde{q}_s$ of q and $\widetilde{q}$, respectively, on $(0, 2\pi)$. It follows that q_s and $\widetilde{q}_s$ have a common Dirichlet spectrum on $(0, 2\pi)$, which is equal to $\sigma_D \cup \sigma_{DN}$. Thus Theorem 2 implies that $q_s(x) \equiv \widetilde{q}_s(x)$ on $[0, 2\pi]$ and therefore $q(x) \equiv \widetilde{q}(x)$ on $[0, \pi]$.

(d) We believe that our approach applies to operators of the form

$$Lu = \left(-\frac{d}{dx}\right)^{2l} u + qu.$$

We end this section with a small result in the spirit of Ambarzumian.

THEOREM 3. *Consider the* (*Neumann*) *eigenvalue problem on the interval* $(0, b)$*,*

$$Lu = u'''' + q(x)u = \mu u, \qquad u'(0) = u'''(0) = u'(b) = u'''(b) = 0$$

with eigenvalues μ_n*,* $n = 0, 1, 2, 3, \ldots$*. If* $q \in C[0, b]$ *and* $\mu_n = n^4$ *for infinitely many* n*'s, including* $n = 0$*, then* $b = \pi$ *and* $q(x) \equiv 0$*.*

Proof. First, we observe that the given behavior of μ_n for large n implies that $b = \pi$ and

$$\int_0^\pi q(x)dx = 0.$$

Next, we consider the associated Rayleigh quotient, namely,

$$N[v] = \frac{\int_0^\pi \left[v''(x)^2 + q(x)v(x)^2\right] dx}{\int_0^\pi v(x)^2 dx}, \qquad v'(0) = v'''(0) = v'(\pi) = v'''(\pi) = 0.$$

We have that $\inf_v N[v] = \mu_0 = 0$. However, $N[1] = 0$. Therefore, $\inf_v N[v] = N[1]$. Thus $v(x) \equiv 1$ is a Neumann eigenfunction corresponding to the eigenvalue 0. Therefore, $q(x) \equiv 0$. □

Appendix. In the second-order case, there is a way to reconstruct q (not mentioned in our introduction) that is due to Hald [11]. Hald's algorithm is a beautiful direct method of solving the inverse problem with symmetric q by a clever discretization. It is suitable for numerical implementation since it gives a stable algorithm for reconstructing q. However, it requires that $||q||_2$ and $\sum_n \left(\lambda_n - n^2\right)^2$ be sufficiently small, and for this reason, its theoretical significance is underestimated since other methods do not impose such restrictions. We believe that Hald's algorithm extends to the fourth-order problem (1).

Here is a telegraphic description of our proposed reconstruction process. First, we mention that if $q \in C^2[0, \pi]$, then the large eigenvalues of (1) obey the asymptotic formula

$$\lambda_n = n^4 + O(n^{-2})$$

(see, for example, [18]). We need to impose the condition

$$\text{(C)} \qquad ||q||_2 \leq Q_2 \quad \text{and} \quad \sum_n \left(\lambda_n - n^4\right)^2 \leq \Lambda_2,$$

where Q_2 and Λ_2 are fixed numerical bounds. (We believe that one can allow bounds that are bigger than the corresponding bounds of [11] for the second-order case.)

Step 1. Let $\lambda_1 < \cdots < \lambda_m$ be the first m eigenvalues of (1). (Condition (C) excludes the possibility of multiple eigenvalues.) Then there is a unique set of real numbers $\beta_1, \dots, \beta_m$ such that if we set

$$b_{jk} = \begin{cases} j^4 \delta_{jk} + \beta_{|k-j|/2} - \beta_{(k+j)/2} & \text{if } k+j = \text{even}, \\ 0 & \text{if } k+j = \text{odd}, \end{cases} \qquad \text{where } \beta_0 = 0,$$

then the $m \times m$ matrix $B = (b_{jk})$ has eigenvalues $\lambda_1, \dots, \lambda_m$ (notice that these β_k's depend on m and that they are not the cosine Fourier coefficients of q), and furthermore,

$$\beta_1^2 + \beta_2^2 + \cdots + \beta_m^2 \leq (2/\pi)\, Q_2^2.$$

Step 2. If we set

$$q_m(x) = 2 \sum_{k=1}^{m} \beta_k \cos(2kx),$$

where β_k, $k = 1, \dots, m$, is as in Step 1, then

$$||q - q_m||_2 \to 0 \quad \text{as } m \to \infty.$$

To justify these steps, we believe that one need only repeat Hald's analysis (see [11]) with basically one modification, namely, the discretization of $-(d/dx)^2$; i.e., the matrix $\operatorname{diag}(1^2, 2^2, 3^2, \dots)$ that appears in [11] should now be replaced by the discretization of $(d/dx)^4$, i.e., the matrix $\operatorname{diag}(1^4, 2^4, 3^4, \dots)$.

Acknowledgments. (i) We want to thank Professor P. Kuchment for showing us certain papers by Yu. Karpeshina.

(ii) Recently, we became aware that Professor Peter Perry and his collaborators, Al Schueller and Lester Caudill, are also working (independently of us) on higher-order inverse spectral problems; some of their methods are similar to those used in this paper.

(iii) Finally, we want to thank the referees for various constructive comments and suggestions.

REFERENCES

[1] V. Barcilon, *Explicit solution of the inverse problem for a vibrating string*, J. Math. Anal. Appl., 93 (1983), pp. 222–234.

[2] G. Borg, *Eine Umkehrung der Sturm–Liouvilleschen Eigenwertaufgabe*, Acta Math., 78 (1946), pp. 1–96 (in German).

[3] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, Robert E. Krieger Publishing Company, Malabar, FL, 1987.

[4] W. Craig, *The trace formula for Schrödinger operators on the line*, Comm. Math. Phys., 126 (1989), pp. 379–407.

[5] N. Dunford and J. T. Schwartz, *Linear Operators Part* II: *Spectral Theory*; *Self Adjoint Operators in Hilbert Space*, Wiley Classics Library Edition, John Wiley, New York, 1988.

[6] M. G. Gasymov and B. M. Levitan, *Determination of a differential equation by two spectra*, Russian Math. Surveys, 19 (1964), pp. 3–63.

[7] I. M. Gelfand and B. M. Levitan, *On the determination of a differential equation from its spectral function*, Izv. Akad. Nauk SSSR Ser. Mat., 15 (1951), pp. 309–360; Amer. Math. Soc. Transl. Ser. 2, 1 (1951), pp. 253–304.

[8] F. Gesztesy, H. Holden, B. Simon, and Z. Zhao, *Trace formulae and inverse spectral theory for Schrödinger operators*, Bull. Amer. Math. Soc. (N. S.), 29 (1993), pp. 250–255.

[9] G. M. L. Gladwell, *Inverse Problems in Vibration*, Martinus Nijhoff, Boston, 1986.

[10] O. H. Hald, *Inverse eigenvalue problems for layered media*,Comm. Pure Appl. Math., 30 (1977), pp. 69–94.

[11] O. H. Hald, *The inverse Sturm–Liouville problems and the Rayleigh–Ritz Method*, Math. Comput., 32 (1978), pp. 687–705.

[12] Yu. E. Karpeshina, *Perturbation theory for the Schrödinger operator with a periodic nonsmooth potential*, Soviet Math. Dokl., 40 (1990), pp. 614–618.

[13] H. J. Landau, *The inverse problem for the vocal tract and the moment problem*, SIAM J. Math. Anal., 14 (1983), pp. 1019–1035.

[14] N. Levinson, *The inverse Sturm–Liouville problem*, Mat. Tidsskrift, B (1949), pp. 25–30.

[15] J. R. McLaughlin, *On constructing solutions to an inverse Euler–Bernoulli problem*, in Inverse Problems of Acoustic and Elastic Waves, F. Santosa, et al., eds., SIAM, Philadelphia, PA, 1984, pp. 341–347.

[16] J. R. McLaughlin, *Analytic methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.

[17] M. A. Naimark, *Linear Differential Operators Parts* I *and* II, Frederick Ungar, New York, 1967 and 1968.

[18] V. Papanicolaou, *Trace formulas and the behavior of large eigenvalues*, SIAM J. Math. Anal., 26 (1995), pp. 218–237.

[19] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV: *Analysis of Operators*, Academic Press, New York, 1978.

[20] V. A. Yurko, *Uniqueness of a reconstruction of binomial differential operators from two spectra*, Mat. Zametki, 43 (1988), pp. 356–364 (in Russian); Math. Notes Acad. Sci. USSR, 43 (1988), pp. 205–210 (in English).

[21] V. A. Yurko, *A problem in elasticity theory*, Prikl. Mat. Mekh., 54 (1990), pp. 998–1002 (in Russian); J. Appl. Math. Mech., 54 (1990), pp. 820–824 (in English).

[22] V. A. Yurko, *Recovery of differential operators from the Weyl matrix*, Dokl. Akad. Nauk USSR, 313 (1990) (in Russian); Soviet Math. Dokl., 42 (1991), pp. 229–233 (in English).

[23] W. W. Zachary, *An inverse scattering formalism for higher-order differential operators*, J. Math. Anal. Appl., 117 (1986), pp. 449–495.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 897–902, July 1997

A LEVINSON-TYPE ALGORITHM FOR DISCRETE STATIONARY RANDOM FIELDS*

A. MAKAGON†, C. A. MENSAH†, AND A. G. MIAMEE†

Abstract. A Levinson-type recursive procedure for updating the predictor of a discrete stationary random field is developed. This considerably reduces the number of equations that one must solve in order to recursively update the predictor as more data and information about the process becomes available and thus saves some resources and time. The procedure extends the well-known Levinson recursive algorithm to the case of random fields.

Key words. stationary random field, prediction, Levinson's algorithm

AMS subject classifications. 60G60, 60G25

PII. S0036141095289385

1. Introduction.

Prediction of random fields requires knowledge of the observed past and the covariance of the process. In practice, one must find the coefficients of the predictor and study the error. In general, finding the coefficients involves solving a system of linear equations at each point in time. As more data and information about the process become available, the predictor must be updated. This requires solving an entirely new and larger system of equations, which becomes increasingly difficult and time consuming.

In the case of a stationary sequence $X = \{x(n) : n \in Z\}$, Levinson [6] developed a simple algorithm for computing the predictor $(x(0)|\mathcal{M}_{n+1})$ of $x(0)$ based on $\mathcal{M}_{n+1} = sp\{x(k) : -n-1 \leq k \leq -1\}$, assuming that $(x(0)|\mathcal{M}_n)$ is known. The purpose of this paper is to present a similar recursive procedure for stationary fields.

As remarked by Kallianpur, Miamee, and Niemi [4], the starting point in the prediction problem is the appropriate definition of the past. In one of the earliest studies, Chiang [1] took the half-plane as the past, whereas Helson and Lowdenslager used the augmented half-plane for past in their study [2, 3]. More recently, Kallianpur and Mandrekar [5] performed a time-domain analysis of a two-parameter weakly stationary random field in which they took the quarter-plane in the southwest corner as the past. Marzetta [7] pointed out that the proper definition of the past must include all of the data points that lie between the data points at (s, t) being predicted and data points at (k, l) that have been observed. In this paper, we will work with the truncated quarter-plane, namely $M_n = \{(i, j) \in Z^2 : -n \leq i, j \leq -1\}$, as the finite past. Having identified our past here, we define the "nearest future" as all those data points that are one step ahead of the finite past, that is, $F_n = \{(i, 0) : -n-1 \leq i \leq 0\} \cup \{(0, j) : -n-1 \leq j \leq -1\}$.

Recall that the Levinson's algorithm (the one-parameter case) is based on the fact that the projection $(x(0)|\mathcal{M}_{n+1})$ of $x(0)$ onto $\mathcal{M}_{n+1}$ can be expressed in terms of $(x(0)|\mathcal{M}_n)$ and the "backward" predictor of the "new observation" $(x(-n-1)|\mathcal{M}_n)$, where the latter can be deduced from the coefficients of $(x(0)|\mathcal{M}_n)$. In the case considered in this paper, the "new observations" consist of $2n+1$ points $x(p, q)$,

* Received by the editors July 24, 1995; accepted for publication (in revised form) March 12, 1996. This research was supported by Office of Naval Research grant N00014-89-J-1824 and U.S. Army Research grant DAAH04-96-1-0027.

http://www.siam.org/journals/sima/28-4/28938.html

† Department of Mathematics, Hampton University, Hampton, VA 26668 (makagon@math1.math.hamptonu.edu, miamee@cs.hamptonu.edu).

$(p,q) \in M_{n+1} \setminus M_n$. Therefore, if one wants to employ Levinson's idea of "backward" prediction, the predictors of all elements $x(p,q)$, $(p,q) \in F_n$, must be available. In other words, the procedure must involve the simultaneous construction of the predictors of all elements in the near future.

2. The problem. Let Z denote the set of integers. A family of complex random variables $X = \{x(i,j) : (i,j) \in Z^2\} \subset L^2(\Omega)$ is said to be a *weakly stationary random field* (*WSRF*) if $E(x(i,j)) = 0$ for all $(i,j) \in Z^2$ and the *correlation function* of X, $E(x(i,j)\overline{x(k,m)}) = R(i-k, j-m)$, depends only on $i-k$ and $j-m$. In what follows, the expectation $E(x\overline{y})$, $x, y \in L^2(\Omega)$, will be denoted by $\langle x, y\rangle$ and $(x|\mathcal{M})$ will stand for the orthogonal projection of x onto a closed subspace $\mathcal{M} \subset L^2(\Omega)$ (with respect to the inner product defined above). Let us recall that

$$y = (x|\mathcal{M}) \quad \Longleftrightarrow \quad y \in \mathcal{M} \text{ and } \langle y, z\rangle = \langle x, z\rangle \tag{1}$$

for each z in a linearly dense subset of $\mathcal{M}$.

For every integer $n > 0$, let $M_n = \{(i,j) \in Z^2 : -n \le i \le -1, -n \le j \le -1\}$ be the finite past, $F_n = \{(i,0) : -n-1 \le i \le 0\} \cup \{(0,j) : -n-1 \le j \le -1\}$ be the nearest future, and $D_n = M_{n+1} \setminus M_n$. Let X be a WSRF and let $\mathcal{M}_n = \operatorname{sp}\{x(i,j) : (i,j) \in M_n\}$, where sp denotes the closed linear span. The *best linear predictor of* $x(p,q)$ *based on* $\mathcal{M}_n$ is the orthogonal projection

$$(x(p,q)|\mathcal{M}_n) = \sum_{(i,j)\in M_n} \alpha_{i,j}^{(n)}(p,q)x(i,j). \tag{2}$$

The coefficients $\alpha_{i,j}^{(n)}(p,q)$, $(i,j) \in M_n$, will be referred to as *predictor coefficients* of the predictor $(x(p,q)|\mathcal{M}_n)$. If we assume that any finite nontrivial subset of X is linearly independent (as vectors in $L^2(\Omega)$), then by (1), the predictor coefficients are the unique solution of the following system of linear equations:

$$R(p-k, q-l) = \sum_{(i,j)\in M_n} \alpha_{i,j}^{(n)}(p,q)R(i-k, j-l), \quad (k,l) \in M_n. \tag{3}$$

Therefore, finding the predictor of the "nearest future" elements $\{x(p,q) : (p,q) \in F_n\}$ based on $\mathcal{M}_n$ requires solving $(2n+3)$ systems of linear equations each with n^2 unknowns. In the next section, we present a recursive Levinson-type algorithm for computing the predictor coefficients.

3. The algorithm. In this section, X is a WSRF with the property that any finite nontrivial subset of X is linearly independent. Let $n > 0$ be fixed and assume that the predictor coefficients $\alpha_{i,j}^{(n)}(p,q)$, $(i,j) \in M_n$, of $(x(p,q)|\mathcal{M}_n)$ are known for every $(p,q) \in F_n$. To get our prediction of the nearest future at the next iteration, we have to find the coefficients $\alpha_{i,j}^{(n+1)}(p,q)$ for $(i,j) \in M_{n+1}$ and $(p,q) \in F_{n+1}$. We first find these coefficients for $(p,q) \in F_n$. The remaining two sets of coefficients, namely $\alpha_{i,j}^{(n+1)}(-n-2, 0)$ and $\alpha_{i,j}^{(n+1)}(0, -n-2)$, will be discussed later.

Denote $e_n(s,t) = x(s,t) - (x(s,t)|\mathcal{M}_n)$ and $\mathcal{E}_n = \operatorname{sp}\{e_n(s,t) : (s,t) \in D_n\}$. Then $e_n(s,t), (s,t) \in D_n$, are linearly independent and $\mathcal{E}_n = \mathcal{M}_{n+1} \ominus \mathcal{M}_n$. Hence for every (p,q),

$$(x(p,q)|\mathcal{M}_{n+1}) = (x(p,q)|\mathcal{M}_n) + \sum_{(s,t)\in D_n} \beta_{s,t}^{(n)}(p,q)e_n(s,t). \tag{4}$$

Substituting $e_n(s,t) = x(s,t) - (x(s,t)|\mathcal{M}_n)$ and then using (2), we obtain

$$(x(p,q)|\mathcal{M}_{n+1}) = \sum_{(i,j)\in M_n} \left[\alpha_{i,j}^{(n)}(p,q) - \sum_{(s,t)\in D_n} \beta_{s,t}^{(n)}(p,q)\alpha_{i,j}^{(n)}(s,t) \right] x(i,j) + \sum_{(i,j)\in D_n} \beta_{i,j}^{(n)}(p,q)x(i,j).$$

This implies that

$$\alpha_{i,j}^{(n+1)}(p,q) = \begin{cases} \alpha_{i,j}^{(n)}(p,q) - \sum\limits_{(s,t)\in D_n} \beta_{s,t}^{(n)}(p,q)\alpha_{i,j}^{(n)}(s,t) & \text{if } (i,j)\in M_n, \\ \beta_{i,j}^{(n)}(p,q) & \text{if } (i,j)\in D_n. \end{cases} \tag{5}$$

Relation (5) shows that for any $(p,q)\in F_n$, in order to obtain the predictor coefficients of $(x(p,q)|\mathcal{M}_{n+1})$, it is enough to find the coefficients $\alpha_{i,j}^{(n)}(s,t)$, $(s,t)\in D_n$, $(i,j)\in M_n$, and $\beta_{s,t}^{(n)}(p,q)$, $(s,t)\in D_n$.

In the next lemma, we show that the coefficients $\alpha_{i,j}^{(n)}(s,t)$, $(s,t)\in D_n$, can be obtained by renumbering the coefficients $\alpha_{i,j}^{(n)}(p,q)$, $(p,q)\in F_n$. This is a simple consequence of the fact that due to the stationarity of X, the inner products $\langle x(s,t),\, x(i,j)\rangle$ and $\langle x(-n-1-i,-n-1-j),\, x(-n-1-s,-n-1-t)\rangle$ are equal. This principle, usually called the principle of backward–forward predictors, is the basis of Levinson's algorithm.

LEMMA 1. *For every $(i,j)\in M_n$ and $(s,t)\in D_n$,*

$$\alpha_{i,j}^{(n)}(s,t) = \overline{\alpha_{-n-1-i,-n-1-j}^{(n)}(-n-1-s,-n-1-t)}.$$

Note that if $(s,t)\in D_n$, then $(-n-1-s,-n-1-t)\in F_n$.

Proof. By (3), it suffices to show that for all $(u,v)\in M_n$,

$$R(s-u,t-v) = \sum_{(i,j)\in M_n} \overline{\alpha_{i,j}^{(n)}(-n-1-s,-n-1-t)}R(i-u,j-v). \tag{6}$$

Using the fact that $R(i,j) = \overline{R(-i,-j)}$ and substituting $i=-n-1-k$ and $j=-n-1-l$, we see that the right-hand side of (6) equals

$$\sum_{(k,l)\in M_n} \overline{\alpha_{k,l}^{(n)}(-n-1-s,-n-1-t)R(k-(-n-1-u),l-(-n-1-v))}.$$

Using (3) again, this time with $p=-n-1-s$ and $q=-n-1-t$, we see that the latter equals $R(u-s,v-t)$. □

The next lemma provides the equations for computing the coefficients $\beta_{s,t}^{(n)}(p,q)$, $(s,t)\in D_n$, $(p,q)\in F_n$, defined in (4). We state the equations in a more general form that is ready to be used later in computing the predictor coefficients of $(x(-n-2,0)|\mathcal{M}_{n+1})$. Let us remark that in the case of a one-parameter stationary sequence, the space $\mathcal{E}_n$ is one dimensional and there is only one coefficient to compute.

LEMMA 2. *Let L_n be the "nearest neighborhood" of M_n, that is, $L_n = F_n\cup D_n$, and let K be a nonempty subset of L_n. Let $\mathcal{K} = \mathrm{sp}\{x(i,j) : (i,j)\in M_n\cup K\}\ominus\mathcal{M}_n$.*

Denote

$$\begin{aligned} A(s,t;u,v) &= R(s-u,t-v) - \sum_{(i,j)\in M_n} \alpha_{i,j}^{(n)}(s,t)R(i-u,j-v) \\ &\quad - \sum_{(i,j)\in M_n} \overline{\alpha_{i,j}^{(n)}(u,v)}R(s-i,t-j) \\ &\quad + \sum_{(i,j)\in M_n}\sum_{(k,l)\in M_n} \alpha_{i,j}^{(n)}(s,t)\overline{\alpha_{k,l}^{(n)}(u,v)}R(i-k,j-l), \\ B(p,q;u,v) &= R(p-u,q-v) - \sum_{(i,j)\in M_n} \overline{\alpha_{i,j}^{(n)}(u,v)}R(p-i,q-j). \end{aligned}$$

Then for every $(p,q)\in F_n$,

$$(x(p,q)|\mathcal{M}_n\oplus\mathcal{K}) = (x(p,q)|\mathcal{M}_n) + \sum_{(s,t)\in K} \kappa_{s,t}^{(n)}(p,q)(x(s,t)-(x(s,t)|\mathcal{M}_n)), \tag{7}$$

where the coefficients $\kappa_{s,t}^{(n)}(p,q)$, $(s,t)\in K$, *satisfy the following system of equations:*

$$\sum_{(s,t)\in K} \kappa_{s,t}^{(n)}(p,q)\ A(s,t;u,v) = B(p,q;u,v), \quad (u,v)\in K. \tag{8}$$

Proof. The summation in (7) represents the orthogonal projection on $\mathcal{K}$, and therefore by (1) the coefficients $\kappa_{s,t}^{(n)}(p,q)$, $(s,t)\in K$, are the unique solutions of the system

$$\begin{aligned} \sum_{(s,t)\in K} \kappa_{s,t}^{(n)}(p,q)\langle(x(s,t)-(x(s,t)|\mathcal{M}_n),(x(u,v)-(x(u,v)|\mathcal{M}_n))\rangle \\ = \langle x(p,q),(x(u,v)-(x(u,v)|\mathcal{M}_n))\rangle, \quad (u,v)\in K. \end{aligned}$$

Substituting (2) into the last equation completes the proof. □

If $K=D_n$, Lemma 2 shows that the coefficients $\beta_{s,t}^{(n)}(p,q)$, $(s,t)\in D_n$, $(p,q)\in F_n$, needed in (5), are the solutions of the system

$$\sum_{(s,t)\in D_n} \beta_{s,t}^{(n)}(p,q)\ A(s,t;u,v) = B(p,q;u,v), \quad (u,v)\in D_n, \tag{9}$$

and hence we can compute $(x(p,q)|\mathcal{M}_{n+1})$ for $(p,q)\in F_n$ via formula (7).

It remains to compute the two predictors $(x(-n-2,0)|\mathcal{M}_{n+1})$ and $(x(0,-n-2)|\mathcal{M}_{n+1})$. If we apply Lemma 2 for $K=R_n := \{(0,j) : j=-1,\dots,-n-1\}\cup\{(i,-n-1) : i=-n,\dots,-1\}$ and denote the solutions to (8) by $\rho_{s,t}^{(n)}(p,q)$, we obtain

$$\begin{aligned} &(x(-n-1,0)|\mathcal{M}_n\oplus\mathcal{R}_n) \\ &\quad = (x(-n-1,0)|\mathcal{M}_n) + \sum_{(s,t)\in R_n} \rho_{s,t}^{(n)}(-n-1,0)(x(s,t)-(x(s,t)|\mathcal{M}_n)), \end{aligned} \tag{10}$$

where the coefficients $\rho_{s,t}^{(n)}(-n-1,0)$, $(s,t)\in R_n$, satisfy

$$\sum_{(s,t)\in R_n} \rho_{s,t}^{(n)}(-n-1,0)\ A(s,t;u,v) = B(-n-1,0;u,v), \quad (u,v)\in R_n. \tag{11}$$

The coefficients $\rho_{s,t}^{(n)}$ are needed to compute the coefficients $\alpha_{i,j}^{(n+1)}(-n-2,0)$ and $\alpha_{i,j}^{(n+1)}(0,-n-2)$, $(i,j) \in M_{n+1}$, and this is done in the next lemma. In fact, from the remark preceeding Lemma 1, it follows that it is enough to compute only one of these two sets of coefficients.

Lemma 3. *Let $\rho_{s,t}^{(n)}(-n-1,0)$, $(s,t) \in R_n$, be as above.*

1. *If $(i,j) \in \{-n-1, \ldots, -2\} \times \{-n, \ldots, -1\}$, then*

$$\alpha_{i,j}^{(n+1)}(-n-2,0) = \alpha_{i+1,j}^{(n)}(-n-1,0) + \sum_{(s,t)\in R_n} \rho_{s,t}^{(n)}(-n-1,0)\alpha_{i+1,j}^{(n)}(-n-1,0).$$

2. *If $(i,j) \in \{(-1,j) : -n-1 \le j \le -1\} \cup \{(i,-1-n) : -n-1 \le i \le -2\}$, then*

$$\alpha_{i,j}^{(n+1)}(-n-2,0) = \rho_{i+1,j}^{(n)}(-n-1,0).$$

3. *If $(i,j) \in M_{n+1}$, then*

$$\alpha_{i,j}^{(n+1)}(0,-n-2) = \overline{\alpha_{-n-2-i,-n-2-j}^{(n+1)}(-n-2,0)}.$$

Proof. Consider the unitary operator U defined by $Ux(i,j) = x(i-1,j)$, $(i,j) \in Z^2$. Then by (10), $(x(-n-2,0)|\mathcal{M}_{n+1}) = U(x(-n-1,0)|\mathcal{M}_n \oplus \mathcal{R}_n) = U((x(-n-1,0)|\mathcal{M}_n) + \sum_{(s,t)\in R_n} \rho_{s,t}^{(n)}(-n-1,0)(x(s,t) - (x(s,t)|\mathcal{M}_n))$. Substituting (2), we obtain parts 1 and 2. Part 3 of the lemma follows immediately from the fact that $\langle x(-n-2,0), x(i,j)\rangle = \langle x(0,-n-2), x(n+2+i,-n-2+j)\rangle$ (cf. Lemma 1). □

The following theorem summarizes the algorithm presented in this note.

Theorem 1 (Levinson algorithm for random fields). *Let X be a WSRF with the property that any finite nontrivial subset of X is linearly independent. Assume that the predictor coefficients $\alpha_{i,j}^{(n)}(p,q)$, $(i,j) \in M_n$, of $(x(p,q)|\mathcal{M}_n)$ are known for each $(p,q) \in F_n$. Let the functions $A(\cdot,\cdot\,;\,\cdot,\cdot)$ and $B(\cdot,\cdot\,;\,\cdot,\cdot)$ be as in Lemma 2. Then for every $(p,q) \in F_{n+1}$, the coefficients $\alpha_{i,j}^{(n+1)}(p,q)$, $(i,j) \in M_{n+1}$, appearing in*

$$(x(p,q)|\mathcal{M}_{n+1}) = \sum_{(i,j)\in M_{n+1}} \alpha_{i,j}^{(n+1)}(p,q)x(i,j)$$

can be computed via the following procedure:

(i) *If $(p,q) \in F_{n+1} \setminus \{(-n-2,0),(0,-n-2)\}$, then $\alpha_{i,j}^{(n+1)}(p,q)$ are given by (5), where $\alpha_{i,j}^{(n)}(s,t)$, $(s,t) \in D_n$, are related to $\alpha_{i,j}^{(n)}(p,q)$, $(p,q) \in F_n$, by Lemma 1 and $\beta_{s,t}^{(n)}(p,q)$, $(s,t) \in D_n$, are solutions of the system*

$$\sum_{(s,t)\in D_n} \beta_{s,t}^{(n)}(p,q)\; A(s,t;u,v) = B(p,q;u,v), \quad (u,v) \in D_n. \tag{12}$$

(ii) *If $(p,q) \in \{(-n-2,0),(0,-n-2)\}$, then the coefficients $\alpha_{i,j}^{(n+1)}(p,q)$ are calculated as in Lemma 3.*

From Theorem 1, it follows that in order to find $\alpha_{i,j}^{(n+1)}(p,q)$, $(i,j) \in M_{n+1}$, for a fixed $(p,q) \in F_{n+1}$, the algorithm requires solving system (12) (or (11)) consisting of $2n+1$ equations whose coefficients are obtained by renumbering, multiplying, and summing known predictor coefficients $\alpha_{i,j}^{(n)}(u,v)$, $(u,v) \in F_n$, $(i,j) \in M_n$, while direct computation would involve solving system (3) (with n replaced by $n+1$) with

$(n+1)^2$ equations. In the case of stationary sequences, the original Levinson algorithm required computing only one new coefficient, and this is because in that case there is just one "new observation," while in our case there are $2n+1$ "new observations."

REFERENCES

[1] T. P. Chiang, *On the linear extrapolation of a continuous homogeneous random field,* Theory Probab. Appl., 2 (1957), pp. 58–89.

[2] H. Helson and D. Lowdenslager, *Prediction theory and Fourier series and several variables* I, Acta Math., 99 (1958), pp. 165–202.

[3] H. Helson and D. Lowdenslager, *Prediction theory and Fourier series in several variables* II, Acta Math., 106 (1961), pp. 173–213.

[4] G. Kallianpur, A. G. Miamee, and H. Niemi, *On the prediction of two parameter random fields,* J. Multivariate Anal., 32 (1990), pp. 120–149.

[5] G. Kallianpur and V. Mandrekar, *Non deterministic random fields and Wold and Halmos decomposition for commuting isometries,* in Prediction Theory and Harmonic Analysis: The Pesi Masani Volume, V. Mandrekar and H. Salehi, eds., North–Holland, Amsterdam, 1983, pp. 165–190.

[6] N. Levinson, *The Wiener RMS error criterion in filter design and prediction,* J. Math. Phys., 25 (1946), pp. 261–278.

[7] T. Marzetta, *A linear prediction approach to two dimensional spectral factorization and spectral estimation,* Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1978.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 903–923, July 1997

008

THE GENERAL ANALYTIC SOLUTION OF A FUNCTIONAL EQUATION OF ADDITION TYPE[*]

H. W. BRADEN[†] AND V. M. BUCHSTABER[‡]

Abstract. The general analytic solution to the functional equation

$$\phi_1(x+y) = \frac{\begin{vmatrix} \phi_2(x) & \phi_2(y) \\ \phi_3(x) & \phi_3(y) \end{vmatrix}}{\begin{vmatrix} \phi_4(x) & \phi_4(y) \\ \phi_5(x) & \phi_5(y) \end{vmatrix}}$$

is characterized. Up to the action of the symmetry group, this is described in terms of Weierstrass elliptic functions. We illustrate our theory by applying it to the classical addition theorems of the Jacobi elliptic functions and the functional equations

$$\phi_1(x+y) = \phi_4(x)\phi_5(y) + \phi_4(y)\phi_5(x)$$

and

$$\Psi_1(x+y) = \Psi_2(x+y)\phi_2(x)\phi_3(y) + \Psi_3(x+y)\phi_4(x)\phi_5(y).$$

Key words. functional equation, Calogero–Moser, special functions

AMS subject classifications. 39B32, 33E05

PII. S0036141095291385

1. Introduction.

The purpose of this article is to describe the general analytic solution to the functional equation

$$\phi_1(x+y) = \frac{\begin{vmatrix} \phi_2(x) & \phi_2(y) \\ \phi_3(x) & \phi_3(y) \end{vmatrix}}{\begin{vmatrix} \phi_4(x) & \phi_4(y) \\ \phi_5(x) & \phi_5(y) \end{vmatrix}}. \tag{1.1}$$

Although this equation appears to depend on five a priori unknown functions, we shall show that (1.1) is invariant under a large group of symmetries $\mathcal{G}$ and that each orbit has a solution of a particularly nice form, expressible in terms of elliptic functions.

THEOREM 1. *The general analytic solution to the functional equation* (1.1) *is, up to a* $\mathcal{G}$*action given by* (2.1)–(2.4) (see section 2)*, of the form*

$$\phi_1(x) = \frac{\Phi(x;\nu_1)}{\Phi(x;\nu_2)}, \qquad \begin{pmatrix} \phi_2(x) \\ \phi_3(x) \end{pmatrix} = \begin{pmatrix} \Phi(x;\nu_1) \\ \Phi(x;\nu_1)' \end{pmatrix}, \quad \textit{and} \quad \begin{pmatrix} \phi_4(x) \\ \phi_5(x) \end{pmatrix} = \begin{pmatrix} \Phi(x;\nu_2) \\ \Phi(x;\nu_2)' \end{pmatrix}.$$

Here

$$\Phi(x;\nu) \equiv \frac{\sigma(\nu - x)}{\sigma(\nu)\sigma(x)}\, e^{\zeta(\nu)x}, \tag{1.2}$$

[*] Received by the editors September 5, 1995; accepted for publication March 5, 1996.
http://www.siam.org/journals/sima/28-4/29138.html

[†] Department of Mathematics and Statistics, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom (hwb@maths.ed.ac.uk).

[‡] National Scientific and Research Institute of Physico-Technical and Radio-Technical Measurement (VNIIFTRI), Mendeleevo, Moscow Region 141570, Russia (buchstab@tomogr.msk.su). The research of this author was supported by a 1993 Royal Society Kapitza Fellowship and a 1994/1995 EPSRC Visiting Fellowship.

where $\sigma(x) = \sigma(x|\omega, \omega')$ *and* $\zeta(x) = \sigma(x)'/\sigma(x)$ *are the Weierstrass sigma and zeta functions.*

The group $\mathcal{G}$ of symmetries of (1.1) will be described further below. Our proof is constructive and indeed yields more.

Theorem 2. *Let* x_0 *be a generic point for* (1.1). *Then for* $k = 1, 2$, *we have*

$$\partial_y \ln \left| \begin{matrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{matrix} \right| \Bigg|_{y=0} = \zeta(\nu_k) - \zeta(x) - \zeta(\nu_k - x) - \lambda_k$$

$$= -\frac{1}{x} - \lambda_k + \sum_{l=0} F_l \, \frac{x^{l+1}}{(l+1)!},$$

and the Laurent expansion determines the parameters g_1 *and* g_2 *(which are the same for both* $k = 1, 2$*) characterizing the elliptic functions of* (1.2) *by*

$$g_2 = \frac{5}{3}\left(F_2 + 6F_0^2\right), \qquad g_3 = 6F_0^3 - F_1^2 + \frac{5}{3} F_0 F_2$$

and the parameters ν_k *via* $F_0 = -\wp(\nu_k)$. *Further, we have*

$$\phi_1(x + 2x_0) = \phi_1(2x_0)\, e^{(\lambda_2 - \lambda_1)x} \, \frac{\Phi(x;\nu_1)}{\Phi(x;\nu_2)}$$

and

$$\begin{pmatrix} \phi_{2k}(x+x_0) \\ \phi_{2k+1}(x+x_0) \end{pmatrix} = \frac{e^{-\lambda_k x}}{f(x)} \begin{pmatrix} \phi'_{2k}(x_0) & \phi_{2k}(x_0) \\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \lambda_k & -1 \end{pmatrix} \begin{pmatrix} \Phi(x;\nu_k) \\ \Phi'(x;\nu_k) \end{pmatrix}.$$

Here the function

$$f(x) = \frac{e^{-\lambda_k x}}{\Phi(x;\nu_k)} \; \frac{\left| \begin{matrix} \phi'_{2k}(x_0) & \phi_{2k}(x_0) \\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0) \end{matrix} \right|}{\left| \begin{matrix} \phi_{2k}(x+x_0) & \phi_{2k}(x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(x_0) \end{matrix} \right|}$$

is, in fact, the same for $k = 1, 2$.

The term "generic" will be defined below, and we will give more expressions for the quantities appearing in the theorem.

One merit of writing (1.1) in this general form is that several different functional equations may now be seen as different points on a $\mathcal{G}$ orbit of (1.1). Thus, for example,

$$\phi_1(x+y) = \phi_1(x)\phi_1(y), \tag{1.3}$$

$$\phi_1(x+y) = \phi_4(x)\phi_5(y) + \phi_4(y)\phi_5(x), \tag{1.4}$$

$$A(x+y)[B(x) - B(y)] = A(x)A'(y) - A(y)A'(x) \tag{1.5}$$

are particular[1] examples of (1.1). The functional equation for the exponential (1.3) corresponds to $\nu_1 = \nu_2$ in our solution and the exponential comes wholly from $\mathcal{G}$.

[1] These correspond to
(a) $\phi_2(x) = \phi_1(x)\phi_4(x)$ and $\phi_3(x) = \phi_1(x)\phi_5(x)$,
(b) $\phi_2(x) = \phi_4^2(x)$ and $\phi_3(x) = \phi_5^2(x)$, and
(c) $\phi_1(x) = \phi_2(x) = A(x)$, $\phi_3(x) = A'(x)$, $\phi_4(x) = B(x)$, and $\phi_5(x) = 1$.

Particular cases of (1.4) have been studied in [8] and we shall determine (see Lemma 4) the general solution to (1.4) as an application of our work.

More interesting is equation (1.5), which has been studied by various authors with assumptions of evenness/oddness on the functions appearing [12, 18, 19] or assumptions on the nature of B [16]. The general solution [4, 5] $A(x) = \Phi(x;\nu)$ now corresponds to the limit $\nu_2 \to 0$ together with a $\mathcal{G}$ action. This will be illustrated later.

Finally, when $\phi_1(x) = \alpha(x)$, $\phi_2(x) = \alpha(x)\tau(x)$, $\phi_4(x) = \tau(x)$, $\phi_3(x) = \phi_2'(x)$, and $\phi_5(x) = \phi_4'(x)$, we obtain the functional equation

$$\alpha(x+y) = \alpha(x)\alpha(y) + \tau(x)\tau(y)\psi(x+y). \tag{1.6}$$

The function $\psi(x)$ will be described in more detail in what follows. This equation was studied by Bruschi and Calogero [4] and will be used in our analysis.

Let us remark that both (1.4) and (1.6) may be viewed as limiting cases of the functional equation

$$\Psi_1(x+y) = \Psi_2(x+y)\phi_2(x)\phi_3(y) + \Psi_3(x+y)\phi_4(x)\phi_5(y), \tag{1.7}$$

which a priori depends on seven unknown functions. Later we shall show how (1.1) may be used to solve this.

It remains to place (1.1) in some form of context. The last decade has seen a remarkable confluence of ideas from completely integrable systems, geometry, field theory, and functional equations that is still being assimilated. To make some of these matters concrete, let us consider how such functional equations arise in the context of integrable systems of particles on the line. A pair of matrices L and M such that $\dot{L} = [L, M]$ is known as a Lax pair; this is a zero-curvature condition. Starting with an ansatz for the matrices L and M, one seeks restrictions necessary to obtain equations of motion of some desired form. These restrictions typically involve the study of functional equations. The paradigm for this approach is the Calogero–Moser system [11]. Beginning with the ansatz (for $n \times n$ matrices)

$$\begin{aligned} L_{jk} &= p_j\delta_{jk} + g\,(1-\delta_{jk})A(q_j - q_k), \\ M_{jk} &= g\left[\delta_{jk}\sum_{l\neq j} B(q_j - q_l) - (1-\delta_{jk})C(q_j - q_k)\right], \end{aligned}$$

one finds that $\dot{L} = [L, M]$ yields the equations of motion for the Hamiltonian system ($n \geq 3$)

$$H = \frac{1}{2}\sum_j p_j^2 + g^2 \sum_{j<k} U(q_j - q_k), \quad U(x) = A(x)A(-x) + \text{constant},$$

provided that $C(x) = -A'(x)$ and that $A(x)$ and $B(x)$ satisfy the functional equation (1.5). With this ansatz and assuming $B(x)$ to be even,[2] Calogero [12] found $A(x)$ to be given by (1.2). In this case, the corresponding potential is the Weierstrass $\wp$-function: $A(x)A(-x) = \wp(\nu) - \wp(x)$. The functional equation (1.6) is associated with a different ansatz and yields the relativistic Calogero–Moser systems [3, 21]. Similarly,

[2] This assumption can, in fact, be removed [2].

(1.1) arises from a more general ansatz [2] associated with equations of motion of the form

$$\ddot{q}_j = \sum_{k\neq j}(a + b\dot{q}_j)(a + b\dot{q}_k)V_{jk}(q_j - q_k),$$

which combines both relativistic ($b \neq 0$) and nonrelativistic ($b = 0$) systems together with potentials that can vary between particle pairs. This unifies, for example, Calogero–Moser and Toda systems [20, 21, 22]. The relativistic examples yield the functional equation (1.1), while the nonrelativistic situation involves the functional equation

$$\phi_6(x+y) = \phi_1(x+y)(\phi_4(x) - \phi_5(y)) + \begin{vmatrix} \phi_2(x) & \phi_3(y) \\ \phi_2'(x) & \phi_3'(y) \end{vmatrix}. \tag{1.8}$$

The general analytic solution to (1.8) has yet to be determined, although particular solutions are known. We remark that (1.1) and, after suitable symmetrizing, (1.8) are particular cases of the functional equation

$$\sum_{i=0}^{N} \phi_{3i}(x+y)\begin{vmatrix} \phi_{3i+1}(x) & \phi_{3i+1}(y) \\ \phi_{3i+2}(x) & \phi_{3i+2}(y) \end{vmatrix} = 0, \tag{1.9}$$

with $N = 1$ in the former case and $N = 2$ in the latter. When $\phi_{3i+2} = \phi_{3i+1}'$, Buchstaber and Krichever have discussed (1.9) in connection with functional equations satisfied by Baker–Akhiezer functions [10].

Lax pairs are only one way in which functional equations are associated with integrable systems, and we mention [7, 13, 9, 14] for others. There also appears to be a close connection between these functional equations and the elliptic genera associated with the string-inspired Witten index [15, 17]. Krichever, for example, used the functional equation (1.5) in his proof of the "rigidity" property of elliptic genera [17], and it also appears when discussing rational and pole solutions of the Kadomtsev–Petviashvili (KP) and Korteweg–deVries (KdV) equations [16, 1]. We feel that this connection between functional equations and completely integrable systems is part of a broader and less well-understood aspect of the subject that deserves further attention.

An outline of the paper is as follows. First, we will discuss the group of symmetries of (1.1). These will be used in the proof of Theorem 1. Before turning to the proof, we show in section 3 how the indicated solution indeed satisfies (1.1), using this as a vehicle to recall some of the properties of elliptic functions that we will need throughout. Section 4 is devoted to the proof of Theorem 1 and section 5 to that of Theorem 2. Several applications of our theorems, including the general analytic solution to (1.4) and a discussion of (1.7), are then given in section 6. An appendix is given that contains various elliptic function formulas that we shall make use of.

Various versions of Theorem 1 have appeared in unpublished preprints. In [6], the form of $\phi_1(x)$ only was stated. In [2], we introduced the $\mathcal{G}$ action to give Theorem 1 in its present form. In improving the proof of this, we obtained Theorem 2, given here alongside the better proof of Theorem 1.

2. The group of symmetries.

We next describe the group $\mathcal{G}$ of invariances of (1.1). Theorem 1 gives a representative of each $\mathcal{G}$ orbit on the solutions of (1.1) with a particularly nice form. First, observe that a large group of symmetries $\mathcal{G}$ act on the solutions of (1.1). The transformation

(2.1)
$$\left(\phi_1(x), \begin{pmatrix}\phi_2(x)\\ \phi_3(x)\end{pmatrix}, \begin{pmatrix}\phi_4(x)\\ \phi_5(x)\end{pmatrix}\right) \to \left(c\, e^{\lambda x}\phi_1(x), U\begin{pmatrix}e^{-\lambda' x}\phi_2(x)\\ e^{-\lambda' x}\phi_3(x)\end{pmatrix}, V\begin{pmatrix}e^{\lambda'' x}\phi_4(x)\\ e^{\lambda'' x}\phi_5(x)\end{pmatrix}\right)$$

clearly preserves (1.1), provided that

$$\lambda + \lambda' + \lambda'' = 0, \qquad U, V \in GL_2, \quad \text{and} \quad \det U = c \det V. \tag{2.2}$$

Further, (1.1) is also preserved by

$$\left(\phi_1(x), \begin{pmatrix}\phi_2(x)\\ \phi_3(x)\end{pmatrix}, \begin{pmatrix}\phi_4(x)\\ \phi_5(x)\end{pmatrix}\right) \to \left(\frac{1}{\phi_1(x)}, \begin{pmatrix}\phi_4(x)\\ \phi_5(x)\end{pmatrix}, \begin{pmatrix}\phi_2(x)\\ \phi_3(x)\end{pmatrix}\right) \tag{2.3}$$

and

$$\left(\phi_1(x), \begin{pmatrix}\phi_2(x)\\ \phi_3(x)\end{pmatrix}, \begin{pmatrix}\phi_4(x)\\ \phi_5(x)\end{pmatrix}\right) \to \left(\phi_1(x), f(x)\begin{pmatrix}\phi_2(x)\\ \phi_3(x)\end{pmatrix}, f(x)\begin{pmatrix}\phi_4(x)\\ \phi_5(x)\end{pmatrix}\right). \tag{2.4}$$

We will use these symmetries in our proof of Theorem 1 to find a solution of (1.1) on each $\mathcal{G}$ orbit with a particularly nice form.

3. Illustration of the solution. Before proceeding to the proof, it is instructive to see how the stated solution satisfies (1.1). This will also allow us to introduce some elliptic function identities needed throughout. From the definition of the zeta function, we have

$$\big(\ln\Phi(x;\nu)\big)' = -\zeta(\nu - x) - \zeta(x) + \zeta(\nu). \tag{3.1}$$

Thus

$$\begin{aligned}\begin{vmatrix}\Phi(x;\nu) & \Phi(y;\nu)\\ \Phi(x;\nu)' & \Phi(y;\nu)'\end{vmatrix} &= \Phi(x;\nu)\Phi(y;\nu)\Big[\big(\ln\Phi(y;\nu)\big)' - \big(\ln\Phi(x;\nu)\big)'\Big]\\ &= \Phi(x;\nu)\Phi(y;\nu)\Big[\zeta(\nu - x) + \zeta(x) + \zeta(-y) + \zeta(y-\nu)\Big].\end{aligned}$$

Upon using the definition of Φ, the right-hand side of this equation takes the form

$$\Phi(x+y;\nu)\frac{\sigma(\nu - x)\sigma(\nu - y)\sigma(x+y)}{\sigma(\nu - x - y)\sigma(\nu)\sigma(x)\sigma(y)}\Big[\zeta(\nu - x) + \zeta(x) + \zeta(-y) + \zeta(y-\nu)\Big]. \tag{3.2}$$

After noting the two identities [23]

$$\zeta(x) + \zeta(y) + \zeta(z) - \zeta(x+y+z) = \frac{\sigma(x+y)\sigma(y+z)\sigma(z+x)}{\sigma(x)\sigma(y)\sigma(z)\sigma(x+y+z)} \tag{3.3}$$

and

$$\wp(x) - \wp(y) = \frac{\sigma(y-x)\sigma(y+x)}{\sigma^2(y)\sigma^2(x)}, \tag{3.4}$$

we find that (3.2) simplifies to $\Phi(x+y;\nu)\big[\wp(x)-\wp(y)\big]$, where $\wp(x)=-\zeta'(x)$ is the Weierstrass $\wp$-function. Putting these together yields the addition formula

$$\Phi(x+y;\nu)=\frac{\begin{vmatrix}\Phi(x;\nu) & \Phi(y;\nu)\\ \Phi(x;\nu)' & \Phi(y;\nu)'\end{vmatrix}}{\wp(x)-\wp(y)} \tag{3.5}$$

and consequently a solution of (1.1) with the stated form. Further, from (3.5), we see the solution to (1.5) mentioned in section 1.

The general solution (1.2) involves the two nonzero constants ν_1 and ν_2. Let us see how our group of symmetries enables $\phi_1(x)=\Phi(x;\nu_1)$ to occur as a limit $\nu_2\to 0$. Consider the $\mathcal{G}$ action on the general solution $\phi_i(x)$ of Theorem 1 given by $\phi_i(x)\to\tilde\phi_i(x)$, where

$$\left(\tilde\phi_1(x),\begin{pmatrix}\tilde\phi_2(x)\\ \tilde\phi_3(x)\end{pmatrix},\begin{pmatrix}\tilde\phi_4(x)\\ \tilde\phi_5(x)\end{pmatrix}\right)=\left(\frac{e^{\zeta(\nu_2)x}}{-\nu_2}\phi_1(x),\begin{pmatrix}\Phi(x;\nu_1)\\ \Phi'(x;\nu_1)\end{pmatrix},\begin{pmatrix}e^{-\zeta(\nu_2)x}\Phi(x;\nu_2)\\ -\nu_2\, e^{-\zeta(\nu_2)x}\Phi'(x;\nu_2)\end{pmatrix}\right).$$

Now

$$\lim_{\nu_2\to 0}\tilde\phi_1(x)=\Phi(x;\nu_1)$$

and

$$\lim_{\nu_2\to 0}\begin{vmatrix}\tilde\phi_4(x) & \tilde\phi_4(y)\\ \tilde\phi_5(x) & \tilde\phi_5(y)\end{vmatrix}=\wp(x)-\wp(y).$$

Thus (1.5) arises as the $\nu_2\to 0$ of (1.1).

4. Proof of Theorem 1.

Our proof of Theorem 1 proceeds in two stages. First, we will use the symmetry (2.2) to transform (1.1) into a particularly simple canonical form. This form may be immediately integrated to yield a functional equation studied by Bruschi and Calogero [4]; by appealing to their result, our Theorem 1 will follow. The first stage of this process is entirely algorithmic, and consequently we may readily identify the parameters that appear in our solution. We begin with the following definition.

DEFINITION 1. *A point $x_0\in\mathbb{C}$ is said to be generic for* (1.1) *if*

(1) *$\phi_k(x)$ is regular at x_0 for $k=2,\dots,5$,*

(2) *$\phi_1(x)$ is regular at $2x_0$, and*

(3) $$\begin{vmatrix}\phi_2(x_0) & \phi_2'(x_0)\\ \phi_3(x_0) & \phi_3'(x_0)\end{vmatrix}\neq 0,\qquad \begin{vmatrix}\phi_4(x_0) & \phi_4'(x_0)\\ \phi_5(x_0) & \phi_5'(x_0)\end{vmatrix}\neq 0.$$

Now let x_0 be a generic point. Using at first the matrices U and V of transformation (2.2), we may choose linear combinations of ϕ_k $(k:2,\dots,5)$ such that (1.1) becomes

$$\tilde\phi_1(x+y)=\frac{\begin{vmatrix}\tilde\phi_2(x) & \tilde\phi_2(y)\\ \tilde\phi_3(x) & \tilde\phi_3(y)\end{vmatrix}}{\begin{vmatrix}\tilde\phi_4(x) & \tilde\phi_4(y)\\ \tilde\phi_5(x) & \tilde\phi_5(y)\end{vmatrix}}$$

and such that (for $k=1,2$)

$$\tilde\phi_{2k}(0)=\tilde\phi_{2k+1}'(0)=0,\qquad \tilde\phi_{2k}'(0)=\tilde\phi_{2k+1}(0)=1. \tag{4.1}$$

The arguments of the functions have been shifted to be centered on x_0 (or $2x_0$ in the case of $\tilde{\phi}_1(x)$). Here we have set

$$\tilde{\phi}_1(x) = c\,\phi_1(x + 2x_0)$$

and (for $k = 1, 2$)

$$\begin{pmatrix} \tilde{\phi}_{2k}(x) \\ \tilde{\phi}_{2k+1}(x) \end{pmatrix} = \begin{pmatrix} \phi'_{2k}(x_0) & \phi_{2k}(x_0) \\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0) \end{pmatrix}^{-1} \begin{pmatrix} \phi_{2k}(x + x_0) \\ \phi_{2k+1}(x + x_0) \end{pmatrix}.$$

The constant c here is just the ratio of the appropriate determinants specified in (2.1). We next observe the following result.

Lemma 1. *For $k = 1, 2$, we may write*

$$\begin{pmatrix} \tilde{\phi}_{2k}(x) \\ \tilde{\phi}_{2k+1}(x) \end{pmatrix} = \frac{1}{\gamma_k(x)} \begin{pmatrix} \xi_k(x) \\ \xi'_k(x) \end{pmatrix},$$

where $\gamma_k(x), \xi_k(x)$ are regular at 0 and

$$\xi_k(0) = 0, \qquad \xi'_k(0) = \gamma_k(0) = 1.$$

Further, upon writing $\xi_k(x) = e^{\lambda_k x}\tilde{\xi}_k(x)$ with $\lambda_k = -\tilde{\phi}''_{2k}(0)/2$, the function $\tilde{\xi}_k(x)$, regular at 0, satisfies

$$\tilde{\xi}_k(0) = \tilde{\xi}''_k(0) = 0, \qquad \tilde{\xi}'_k(0) = 1.$$

Proof. Upon differentiating $\xi_k(x) = \gamma_k(x)\tilde{\phi}_{2k}(x)$ and comparing with $\xi'_k(x) = \gamma_k(x)\tilde{\phi}_{2k+1}(x)$, we see that

$$\frac{\gamma'_k(x)}{\gamma_k(x)} = \frac{\tilde{\phi}_{2k+1}(x) - \tilde{\phi}'_{2k}(x)}{\tilde{\phi}_{2k}(x)}. \tag{4.2}$$

The only issue is whether the right-hand side of this differential equation is regular at $x = 0$. Using (4.1) and l'Hôpital's rule, we find that

$$\frac{\gamma'_k(0)}{\gamma_k(0)} = \frac{\tilde{\phi}'_{2k+1}(0) - \tilde{\phi}''_{2k}(0)}{\tilde{\phi}'_{2k}(0)} = -\tilde{\phi}''_{2k}(0),$$

and so $\gamma_k(x)$ and hence $\xi_k(x)$ are regular at 0. Now $\gamma_k(x)$ is determined by (4.2) given an initial condition, which we choose to be $\gamma_k(0) = 1$. The remaining initial conditions for $\xi_k(x)$ follow from (4.1). Indeed, from

$$\tilde{\phi}'_{2k+1}(x) = \frac{\xi''_k(x)\gamma_k(x) - \xi'_k(x)\gamma'_k(x)}{\gamma_k^2(x)},$$

we also find that

$$\xi''_k(0) = -\tilde{\phi}''_{2k}(0).$$

Now upon writing $\xi_k(x) = e^{\lambda_k x}\tilde{\xi}_k(x)$ with $\lambda_k = -\tilde{\phi}''_{2k}(0)/2$, we obtain the final statement of the lemma. □

Thus far we have not used the exponential part of the symmetry (2.2). Utilizing this symmetry, we set $\tilde\xi_0(x) = e^{(\lambda_1-\lambda_2)x}\tilde\phi_1(x)$ and $\gamma(x) = e^{2(\lambda_1-\lambda_2)x}\gamma_2(x)/\gamma_1(x)$. This scaling has (upon noting that $2\lambda_k = \gamma_k'(0)$) the effect of making $\gamma'(0) = 0$. Thus we obtain the following result.

Corollary 1. *At any generic point, we may rewrite* (1.1) *using the symmetry* (2.1) *as*

$$\tilde\xi_0(x+y) = \gamma(x)\gamma(y)\frac{\begin{vmatrix}\tilde\xi_1(x) & \tilde\xi_1(y)\\ \tilde\xi_1'(x) & \tilde\xi_1'(y)\end{vmatrix}}{\begin{vmatrix}\tilde\xi_2(x) & \tilde\xi_2(y)\\ \tilde\xi_2'(x) & \tilde\xi_2'(y)\end{vmatrix}}, \tag{4.3}$$

where for $k = 1, 2$,

$$\tilde\xi_k(0) = \tilde\xi_k''(0) = \gamma'(0) = 0, \qquad \tilde\xi_k'(0) = \gamma(0) = 1. \tag{4.4}$$

Given the complexity of the differential equation (4.2), one may wonder whether (4.3) simplifies much further. In fact, we find the following.

Lemma 2. *The functional equation* (4.3)*—and consequently* (1.1)*—may be written as*

$$\partial\left(\frac{\tilde\xi_1(x+y)}{\tilde\xi_1(x)\tilde\xi_1(y)}\right) = \partial\left(\frac{\tilde\xi_2(x+y)}{\tilde\xi_2(x)\tilde\xi_2(y)}\right), \tag{4.5}$$

where $\partial = \partial_x - \partial_y$. *Further,*

$$\tilde\xi_0(x) = \frac{\tilde\xi_2(x)}{\tilde\xi_1(x)}. \tag{4.6}$$

Proof. Taking the logarithmic derivative of (4.3) with respect to $\partial = \partial_x - \partial_y$, we obtain

$$0 = \frac{\gamma'(x)}{\gamma(x)} - \frac{\gamma'(y)}{\gamma(y)} + \frac{\partial^2\left(\tilde\xi_1(x)\tilde\xi_1(y)\right)}{\partial\left(\tilde\xi_1(x)\tilde\xi_1(y)\right)} - \frac{\partial^2\left(\tilde\xi_2(x)\tilde\xi_2(y)\right)}{\partial\left(\tilde\xi_2(x)\tilde\xi_2(y)\right)}. \tag{4.7}$$

Now employing (4.4), one finds

$$\partial\big(\tilde\xi_k(x)\tilde\xi_k(y)\big)|_{y=0} = \tilde\xi_k'(x)\tilde\xi_k(y) - \tilde\xi_k(x)\tilde\xi_k'(y)|_{y=0} = -\tilde\xi_k(x)$$

and, similarly,

$$\partial^2\big(\tilde\xi_k(x)\tilde\xi_k(y)\big)|_{y=0} = -2\tilde\xi_k'(x).$$

Upon setting $y = 0$ in (4.7) and with these simplifications, we obtain the differential equation

$$0 = \frac{\gamma'(x)}{\gamma(x)} + \frac{2\tilde\xi_1'(x)}{\tilde\xi_1(x)} - \frac{2\tilde\xi_2'(x)}{\tilde\xi_2(x)}$$

with solution

$$\gamma(x) = c\,\frac{\tilde\xi_2^2(x)}{\tilde\xi_1^2(x)}. \tag{4.8}$$

Again using l'Hôpital's rule and (4.4), we find the constant $c = 1$. Therefore, (4.3) may be rewritten as

$$\tilde{\xi}_0(x+y) = \frac{\tilde{\xi}_2^2(x)}{\tilde{\xi}_1^2(x)}\frac{\tilde{\xi}_2^2(y)}{\tilde{\xi}_1^2(y)}\frac{\begin{vmatrix}\tilde{\xi}_1(x) & \tilde{\xi}_1(y)\\ \tilde{\xi}_1'(x) & \tilde{\xi}_1'(y)\end{vmatrix}}{\begin{vmatrix}\tilde{\xi}_2(x) & \tilde{\xi}_2(y)\\ \tilde{\xi}_2'(x) & \tilde{\xi}_2'(y)\end{vmatrix}} = \frac{\begin{vmatrix}\frac{1}{\tilde{\xi}_1(x)} & \frac{1}{\tilde{\xi}_1(y)}\\ \left(\frac{1}{\tilde{\xi}_1(x)}\right)' & \left(\frac{1}{\tilde{\xi}_1(y)}\right)'\end{vmatrix}}{\begin{vmatrix}\frac{1}{\tilde{\xi}_2(x)} & \frac{1}{\tilde{\xi}_2(y)}\\ \left(\frac{1}{\tilde{\xi}_2(x)}\right)' & \left(\frac{1}{\tilde{\xi}_2(y)}\right)'\end{vmatrix}} = \frac{\partial\left(\frac{1}{\tilde{\xi}_1(x)\tilde{\xi}_1(y)}\right)}{\partial\left(\frac{1}{\tilde{\xi}_2(x)\tilde{\xi}_2(y)}\right)}. \tag{4.9}$$

Letting $y \to 0$, we find that

$$\tilde{\xi}_0(x) = \frac{\tilde{\xi}_2(x)}{\tilde{\xi}_1(x)}$$

as required. Utilizing (4.6), we may immediately rewrite (4.9) in the stated form (4.5). □

We observe that at this stage, the symmetry (2.2) has enabled us to transform (1.1) into the form specified by Theorem 1. The solution will follow once we show $1/\tilde{\xi}_k(x) = \Phi(x;\nu_k)$. Now (4.5) may be immediately integrated to give

$$\frac{\tilde{\xi}_1(x+y)}{\tilde{\xi}_1(x)\tilde{\xi}_1(y)} = \frac{\tilde{\xi}_2(x+y)}{\tilde{\xi}_2(x)\tilde{\xi}_2(y)} + \Theta(x+y).$$

Upon setting $\alpha(x) = \tilde{\xi}_2(x)/\tilde{\xi}_1(x)$ and $\psi(x) = \Theta(x)/\tilde{\xi}_2(x)$, this may be rearranged into the form

$$\frac{\alpha(x+y)}{\alpha(x)\alpha(y)} = 1 + \tilde{\xi}_2(x)\tilde{\xi}_2(y)\psi(x+y), \tag{4.10}$$

which is the functional equation studied by Bruschi and Calogero [4]. Calling upon the general analytic solution obtained by these authors together with our initial conditions (4.4), we find that[3] $1/\tilde{\xi}_k(x) = \Phi(x;\nu_k)$ as required.

5. Proof of Theorem 2. It is useful at the outset to gather together the various transformations introduced in the last section:

$$\begin{pmatrix}\tilde{\phi}_{2k}(x)\\ \tilde{\phi}_{2k+1}(x)\end{pmatrix} = \begin{pmatrix}\phi_{2k}'(x_0) & \phi_{2k}(x_0)\\ \phi_{2k+1}'(x_0) & \phi_{2k+1}(x_0)\end{pmatrix}^{-1}\begin{pmatrix}\phi_{2k}(x+x_0)\\ \phi_{2k+1}(x+x_0)\end{pmatrix} \tag{5.1}$$

$$= \frac{1}{\gamma_k(x)}\begin{pmatrix}\xi_k(x)\\ \xi_k'(x)\end{pmatrix} = \frac{e^{\lambda_k x}}{\gamma_k(x)}\begin{pmatrix}1 & 0\\ \lambda_k & 1\end{pmatrix}\begin{pmatrix}\tilde{\xi}_k(x)\\ \tilde{\xi}_k'(x)\end{pmatrix} \tag{5.2}$$

$$= \frac{e^{\lambda_k x}}{\gamma_k(x)\Phi^2(x;\nu_k)}\begin{pmatrix}1 & 0\\ \lambda_k & -1\end{pmatrix}\begin{pmatrix}\Phi(x;\nu_k)\\ \Phi'(x;\nu_k)\end{pmatrix}, \tag{5.3}$$

$$\tilde{\xi}_0(x) = e^{(\lambda_1-\lambda_2)x}\,\frac{\phi_1(x+2x_0)}{\phi_1(2x_0)} = \frac{\Phi(x;\nu_1)}{\Phi(x;\nu_2)}. \tag{5.4}$$

Let us introduce the function

$$f(x) = \gamma_k(x)\Phi^2(x;\nu_k)e^{-2\lambda_k x}. \tag{5.5}$$

[3] For example, from [4], we obtain $\tilde{\xi}_2(x) = Ae^{cx}\sigma(ax|\omega,\omega')/\sigma(ax+\nu|\omega,\omega')$. Using the property $\sigma(ax|a\omega,a\omega') = a\sigma(ax|\omega,\omega')$ and the definition of $\Phi(x;\nu)$, we may rewrite this as $\tilde{\xi}_2(x) = (A/\sigma(\nu/a))e^{(c-\zeta(\nu/a))x}/\Phi(z;-\nu/a)$. Now the $x \to 0$ limit shows $(A/\sigma(\nu/a))e^{(c-\zeta(\nu/a))x} = 1$.

Observe that (4.8) entails that the function $f(x)$ is independent of k:

$$\gamma_1(x)\Phi^2(x;\nu_1)e^{-2\lambda_1 x} = \gamma_2(x)\Phi^2(x;\nu_k)e^{-2\lambda_k x}.$$

With this definition, we may rewrite (5.1) and (5.3) to give

$$\begin{aligned}\begin{pmatrix}\phi_{2k}(x+x_0)\\ \phi_{2k+1}(x+x_0)\end{pmatrix} &= \frac{e^{-\lambda_k x}}{f(x)}\begin{pmatrix}\phi'_{2k}(x_0) & \phi_{2k}(x_0)\\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0)\end{pmatrix}\begin{pmatrix}1 & 0\\ \lambda_k & -1\end{pmatrix}\begin{pmatrix}\Phi(x;\nu_k)\\ \Phi'(x;\nu_k)\end{pmatrix}\\ &= \frac{e^{-\lambda_k x}}{f(x)}\begin{pmatrix}\begin{vmatrix}\Phi(x;\nu_k) & \phi_{2k}(x_0)\\ \Phi'(x;\nu_k) & \phi'_{2k}(x_0)+\lambda_k\phi_{2k}(x_0)\end{vmatrix}\\ \begin{vmatrix}\Phi(x;\nu_k) & \phi_{2k+1}(x_0)\\ \Phi'(x;\nu_k) & \phi'_{2k+1}(x_0)+\lambda_k\phi_{2k+1}(x_0)\end{vmatrix}\end{pmatrix}\end{aligned} \tag{5.6}$$

and

$$\begin{aligned}\begin{pmatrix}\Phi(x;\nu_k)\\ \Phi'(x;\nu_k)\end{pmatrix} &= f(x)e^{\lambda_k x}\begin{pmatrix}1 & 0\\ \lambda_k & -1\end{pmatrix}\begin{pmatrix}\phi'_{2k}(x_0) & \phi_{2k}(x_0)\\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0)\end{pmatrix}^{-1}\begin{pmatrix}\phi_{2k}(x+x_0)\\ \phi_{2k+1}(x+x_0)\end{pmatrix}\\ &= \frac{f(x)e^{\lambda_k x}}{\begin{vmatrix}\phi'_{2k}(x_0) & \phi_{2k}(x_0)\\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0)\end{vmatrix}}\begin{pmatrix}\begin{vmatrix}\phi_{2k}(x+x_0) & \phi_{2k}(x_0)\\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(x_0)\end{vmatrix}\\ \begin{vmatrix}\phi_{2k}(x+x_0) & \phi'_{2k}(x_0)+\lambda_k\phi_{2k}(x_0)\\ \phi_{2k+1}(x+x_0) & \phi'_{2k+1}(x_0)+\lambda_k\phi_{2k+1}(x_0)\end{vmatrix}\end{pmatrix}.\end{aligned} \tag{5.7}$$

Now (5.4) and (5.6) are of the form stated in Theorem 2, provided that we can show that $f(x)$, defined in (5.5), can also be put into the form of the theorem. To see this, note that (5.1) shows

$$\begin{aligned}\tilde{\phi}_{2k}(x) &= \frac{\phi_{2k+1}(x_0)\phi_{2k}(x+x_0)-\phi_{2k}(x_0)\phi_{2k+1}(x+x_0)}{\phi_{2k+1}(x_0)\phi'_{2k}(x_0)-\phi_{2k}(x_0)\phi'_{2k+1}(x_0)}\\ &= \frac{\begin{vmatrix}\phi_{2k}(x+x_0) & \phi_{2k}(x_0)\\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(x_0)\end{vmatrix}}{\begin{vmatrix}\phi'_{2k}(x_0) & \phi_{2k}(x_0)\\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0)\end{vmatrix}},\end{aligned} \tag{5.8}$$

while from (5.3), we see that

$$\gamma_k(x) = \frac{e^{\lambda_k x}}{\Phi(x;\nu_k)\tilde{\phi}_{2k}(x)}. \tag{5.9}$$

Combining these thus shows that

$$f(x) = \frac{e^{-\lambda_k x}}{\Phi(x;\nu_k)}\frac{\begin{vmatrix}\phi'_{2k}(x_0) & \phi_{2k}(x_0)\\ \phi'_{2k+1}(x_0) & \phi_{2k+1}(x_0)\end{vmatrix}}{\begin{vmatrix}\phi_{2k}(x+x_0) & \phi_{2k}(x_0)\\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(x_0)\end{vmatrix}} \tag{5.10}$$

as required. Also, from (5.8) and the definition $\lambda_k = -\tilde{\phi}''_{2k}(0)/2$, we find that

$$-2\lambda_k = \frac{\phi_{2k+1}(x_0)\phi''_{2k}(x_0) - \phi_{2k}(x_0)\phi''_{2k+1}(x_0)}{\phi_{2k+1}(x_0)\phi'_{2k}(x_0) - \phi_{2k}(x_0)\phi'_{2k+1}(x_0)} \tag{5.11}$$

$$= \partial_x \ln \left. \begin{vmatrix} \phi'_{2k}(x+x_0) & \phi_{2k}(x_0) \\ \phi'_{2k+1}(x+x_0) & \phi_{2k+1}(x_0) \end{vmatrix} \right|_{x=0}. \tag{5.12}$$

At this stage, we then see that if we can determine $\Phi(x;\nu_k)$, all of the terms in (5.1)–(5.4) are determined and we obtain the stated expressions for $\phi_1(x)$, $\phi_2(x)$, $\phi_3(x)$, $\phi_4(x)$, $\phi_5(x)$, and $f(x)$ given in Theorem 2. It therefore remains to determine the parameters g_2 and g_3 specifying the elliptic functions $\Phi(x;\nu_k)$ as well as ν_1 and ν_2. To this end, we utilize (5.7) to give

$$\frac{\Phi'(x;\nu_k)}{\Phi(x;\nu_k)} - \lambda_k = \frac{\begin{vmatrix} \phi_{2k}(x+x_0) & \phi'_{2k}(x_0) \\ \phi_{2k+1}(x+x_0) & \phi'_{2k+1}(x_0) \end{vmatrix}}{\begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(x_0) \end{vmatrix}}$$

$$= \partial_y \ln \left. \left| \begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{vmatrix} \right| \right|_{y=0}.$$

Upon using (3.1) to simplify the left-hand side of this equality, we obtain the first equality of Theorem 2,

$$\partial_y \ln \left. \left| \begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{vmatrix} \right| \right|_{y=0} = \zeta(\nu_k) - \zeta(x) - \zeta(\nu_k - x) - \lambda_k, \tag{5.13}$$

and consequently

$$\partial_x\partial_y \ln \left. \left| \begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{vmatrix} \right| \right|_{y=0} = \wp(x) - \wp(\nu_k - x). \tag{5.14}$$

In fact, we have the more general result

$$\begin{aligned} \partial_x\partial_y \ln \begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{vmatrix} &= \partial_x\partial_y \ln \begin{vmatrix} \Phi(x;\nu_k) & \Phi(y;\nu_k) \\ \Phi'(x;\nu_k) & \Phi'(y;\nu_k) \end{vmatrix} \\ &= \partial_x\partial_y \ln[\Phi(x+y;\nu_k)(\wp(x)-\wp(y))] \\ &= \wp(x+y) - \wp(\nu_k - x - y) + \frac{\wp'(x)\wp'(y)}{(\wp(x)-\wp(y))^2}, \end{aligned}$$

from which (5.14) arises as the $y \to 0$ limit.

It remains to show that the Laurent series of (5.13) (or, equivalently, of (5.14)) determines the parameters of $\Phi(x;\nu_k)$. Set

$$\zeta(\nu_k) - \zeta(x) - \zeta(\nu_k - x) - \lambda_k = -\frac{1}{x} - \lambda_k + \sum_{l=0} F_l \, \frac{x^{l+1}}{(l+1)!} \tag{5.15}$$

or, equivalently,

$$\wp(x) - \wp(\nu_k - x) = \frac{1}{x^2} + \sum_{l=0} \frac{F_l}{l!} \, x^l. \tag{5.16}$$

While the coefficients F_l in these expansions depend on $k = 1, 2$, we will avoid including this in our notation; certainly, the combinations of these coefficients that give g_2 and g_3 are independent of k. Now the left-hand side of (5.15) has the expansion

$$-\frac{1}{x} - \lambda_k - \wp(\nu_k)\,x + \wp'(\nu_k)\,\frac{x^2}{2} + (2c_2 - \wp''(\nu_k))\,\frac{x^3}{3!} + \cdots,$$

while that of (5.16) begins with

$$\frac{1}{x^2} + c_2\,x^2 + c_3\,x^4 + \cdots - \left\{\wp(\nu_k) - x\wp'(\nu_k) + \frac{x^2}{2}\wp''(\nu_k) + \cdots\right\}.$$

From either of these, we see that

$$F_0 = -\wp(\nu_k), \qquad F_1 = \wp'(\nu_k), \qquad F_2 = 2c_2 - \wp''(\nu_k),$$

whereupon utilizing (A.4) (see the appendix) we obtain

$$c_2 = \frac{F_2 + 6F_0^2}{12} = \frac{g_2}{20} \quad \text{and} \quad g_3 = 6F_0^3 - F_1^2 + \frac{5}{3}F_0F_2. \tag{5.17}$$

Thus, as stated in Theorem 2, we may obtain the parameters of the elliptic functions from the Laurent expansion (5.15) for either choice of k, the combinations of the coefficients in (5.17) being independent of k. The constant terms in the two expansions then determine ν_1 and ν_2 via $F_0 = -\wp(\nu_k; g_2, g_3)$.

We have now established all of Theorem 2. It is perhaps useful to conclude the section with a lemma that implements the theorem.

LEMMA 3. *Let*

$$\partial_x\partial_y \ln \begin{vmatrix} \phi_{2k}(x+x_0) & \phi_{2k}(y+x_0) \\ \phi_{2k+1}(x+x_0) & \phi_{2k+1}(y+x_0) \end{vmatrix}_{y=0} = -\frac{1}{x} - \lambda_k + \sum_{l=0} F_l\,\frac{x^{l+1}}{(l+1)!}$$
$$= \frac{h'(0)h'(x)}{(h(x)-h(0))^2},$$

where $h(x) = \phi_{2k}(x+x_0)/\phi_{2k+1}(x+x_0)$. *Set* $h_k = h^{(k+1)}(0)/(k+1)!\,h'(0)$. *Then*

$$F_0 = -\begin{vmatrix} h_1 & 1 \\ h_2 & h_1 \end{vmatrix}, \qquad F_1 = 2\begin{vmatrix} h_1 & 1 & 0 \\ h_2 & h_1 & 1 \\ h_3 & h_2 & h_1 \end{vmatrix}, \qquad F_2 = -6\begin{vmatrix} h_1 & 1 & 0 & 0 \\ h_2 & h_1 & 1 & 0 \\ h_3 & h_2 & h_1 & 1 \\ h_4 & h_3 & h_2 & h_1 \end{vmatrix}.$$

6. Examples. We shall now consider the classical addition theorems of the Jacobi elliptic functions and then the functional equations (1.4) and (1.7) as examples of our theory. We have collected several standard results pertaining to elliptic functions that are of use in our computations in the appendix.

Example 1. As a first application of our theory, we consider the addition theorems for the Jacobi elliptic functions $\mathrm{dn}(x)$, $\mathrm{cn}(x)$, and $\mathrm{sn}(x)$, where $\mathrm{dn}(x) \equiv \mathrm{dn}(x|m)$ and so on. These may be cast in the form of (1.1) as

$$\mathrm{dn}(x+y) = \frac{\begin{vmatrix} \mathrm{cn}'(x) & \mathrm{cn}'(y) \\ \mathrm{cn}(x) & \mathrm{cn}(y) \end{vmatrix}}{\begin{vmatrix} \mathrm{sn}'(x) & \mathrm{sn}'(y) \\ \mathrm{sn}(x) & \mathrm{sn}(y) \end{vmatrix}} \quad \text{(Jacobi)}, \qquad \mathrm{cn}(x+y) = \frac{1}{k^2}\frac{\begin{vmatrix} \mathrm{dn}'(x) & \mathrm{dn}'(y) \\ \mathrm{dn}(x) & \mathrm{dn}(y) \end{vmatrix}}{\begin{vmatrix} \mathrm{sn}'(x) & \mathrm{sn}'(y) \\ \mathrm{sn}(x) & \mathrm{sn}(y) \end{vmatrix}},$$

and

$$\operatorname{sn}(x+y) = \frac{\begin{vmatrix} 1 & 1 \\ \operatorname{sn}^2(x) & \operatorname{sn}^2(y) \end{vmatrix}}{\begin{vmatrix} \operatorname{sn}'(x) & \operatorname{sn}'(y) \\ \operatorname{sn}(x) & \operatorname{sn}(y) \end{vmatrix}} \quad \text{(Cayley)}.$$

Let us now apply our theorem to the first equality. The first step is to choose an appropriate generic point x_0. This means that we wish x_0 to be a regular point for $\operatorname{cn}(x)$, $\operatorname{sn}(x)$, and $\operatorname{dn}(2x)$ as well as

$$0 \neq \begin{vmatrix} \operatorname{cn}'(x_0) & \operatorname{cn}''(x_0) \\ \operatorname{cn}(x_0) & \operatorname{cn}'(x_0) \end{vmatrix} = 1 - m + m \operatorname{cn}^4(x_0),$$

$$0 \neq \begin{vmatrix} \operatorname{sn}'(x_0) & \operatorname{sn}''(x_0) \\ \operatorname{sn}(x_0) & \operatorname{sn}'(x_0) \end{vmatrix} = 1 - m \operatorname{sn}^4(x_0).$$

Thus we can take $x_0 = 0$ for this example.

Using (5.11), we find that

$$-2\lambda_1 = \partial_x \ln \operatorname{cn}'(x)\big|_{x=0} = 0 \quad \text{and} \quad -2\lambda_2 = \partial_x \ln \operatorname{sn}'(x)\big|_{x=0} = 0.$$

Further, with $h(x) = \phi_2(x)/\phi_3(x) = \operatorname{cn}'(x)/\operatorname{cn}(x)$, we obtain

$$F(x) = \frac{1 - m + m \operatorname{cn}^4(x)}{\operatorname{sn}^2(x)\operatorname{dn}^2(x)} = \frac{1}{x^2} + \frac{1-2m}{3} + \frac{1 + 14m - 14m^2}{15}x^2 + \cdots,$$

while with $h(x) = \phi_4(x)/\phi_5(x) = \operatorname{sn}'(x)/\operatorname{sn}(x)$, we obtain

$$F(x) = \frac{1 - m \operatorname{sn}^4(x)}{\operatorname{sn}^2(x)} = \frac{1}{x^2} + \frac{1+m}{3} + \frac{1 - 16m + m^2}{15}x^2 + \cdots.$$

In both cases, we find that

$$g_2 = \frac{4}{3}(1 - m + m^2) \quad \text{and} \quad g_3 = \frac{4}{27}(m-2)(2m-1)(m+1)$$

(the required equality providing a nontrivial check), which means that

$$e_1 = \frac{2-m}{3}, \qquad e_2 = \frac{2m-1}{3}, \quad \text{and} \quad e_3 = \frac{-1-m}{3}.$$

Further,

$$\wp(\nu_1) = \frac{2m-1}{3} \quad \text{and} \quad \wp(\nu_2) = \frac{-1-m}{3}.$$

Comparison with (A.7) and (A.9) (see the appendix) shows that $\omega = K(m)$, $\omega' = iK'(m)$, $\nu_1 = K(m) + iK'(m)$, and $\nu_2 = K'(m)$. We may also calculate $f(x) = 1/\operatorname{sn}^2(x)$, and upon using (A.13) (see the appendix), our identity may be rewritten as

$$\operatorname{dn}(x+y) = \frac{\Phi(x+y; K(m) + iK'(m))}{\Phi(x+y; iK'(m))}$$

$$= \frac{\begin{vmatrix} \Phi(x;K(m)+iK'(m)) & \Phi(y;K(m)+iK'(m)) \\ \Phi(x;K(m)+iK'(m))' & \Phi(y;K(m)+iK'(m))' \end{vmatrix}}{\begin{vmatrix} \Phi(x;iK'(m)) & \Phi(y;iK'(m)) \\ \Phi(x;iK'(m))' & \Phi(y;iK'(m))' \end{vmatrix}}$$

$$= \frac{\frac{1}{\mathrm{sn}^2(x)}\begin{vmatrix} \mathrm{cn}'(x) & \mathrm{cn}'(y) \\ \mathrm{cn}(x) & \mathrm{cn}(y) \end{vmatrix}}{\frac{1}{\mathrm{sn}^2(x)}\begin{vmatrix} \mathrm{sn}'(x) & \mathrm{sn}'(y) \\ \mathrm{sn}(x) & \mathrm{sn}(y) \end{vmatrix}}.$$

The second identity may be treated in the same manner, yielding $\nu_1 = K(m)$ and $\nu_2 = K'(m)$. The third identity is a little different. It may be rewritten as

$$\Phi(x+y;iK'(m)) = \frac{1}{\mathrm{sn}(x+y)} = \frac{\frac{1}{\mathrm{sn}^2(x)}\begin{vmatrix} \mathrm{sn}'(x) & \mathrm{sn}'(y) \\ \mathrm{sn}(x) & \mathrm{sn}(y) \end{vmatrix}}{\frac{1}{\mathrm{sn}^2(x)}\begin{vmatrix} 1 & 1 \\ \mathrm{sn}^2(x) & \mathrm{sn}^2(y) \end{vmatrix}}$$

$$= \frac{\begin{vmatrix} \Phi(x;iK'(m)) & \Phi(y;iK'(m)) \\ \Phi(x;iK'(m))' & \Phi(y;iK'(m))' \end{vmatrix}}{\begin{vmatrix} \Phi^2(x;iK'(m)) & \Phi^2(y;iK'(m)) \\ 1 & 1 \end{vmatrix}}.$$

Now

$$\Phi^2(x;iK'(m)) - \Phi^2(y;iK'(m)) = \wp(x) - \wp(y),$$

and the required identity follows from the general solution by the limiting procedure described in section 3.

Example 2. We shall now determine the general analytic solution of

$$\phi_1(x+y) = \phi_4(x)\phi_5(y) + \phi_4(y)\phi_5(x) = \frac{\begin{vmatrix} \phi_2(x) & \phi_2(y) \\ \phi_3(x) & \phi_3(y) \end{vmatrix}}{\begin{vmatrix} \phi_4(x) & \phi_4(y) \\ \phi_5(x) & \phi_5(y) \end{vmatrix}},$$

where $\phi_2(x) = \phi_4^2(x)$ and $\phi_3(x) = \phi_5^2(x)$. The particular case $\phi_1(x) = \phi_4(x)$ was treated in [8].

Suppose that x_0 is a generic point. Then from

$$0 \neq \begin{vmatrix} \phi_2(x_0) & \phi_2'(x_0) \\ \phi_3(x_0) & \phi_3'(x_0) \end{vmatrix} = 2\phi_4(x_0)\phi_5(x_0)\begin{vmatrix} \phi_4(x_0) & \phi_4'(x_0) \\ \phi_5(x_0) & \phi_5'(x_0) \end{vmatrix},$$

we see that $\phi_4(x_0) \neq 0$, $\phi_5(x_0) \neq 0$, and $\phi_1(2x_0) = 2\phi_4(x_0)\phi_5(x_0) \neq 0$. Further, from (5.11), we find

$$\lambda_1 = \lambda_2 - \frac{1}{2}\left(\frac{\phi_4'(x_0)}{\phi_4(x_0)} + \frac{\phi_5'(x_0)}{\phi_5(x_0)}\right). \tag{6.1}$$

Our strategy is as follows. We will first determine ν_1, ν_2, λ_1, and λ_2, the parameters that describe the elliptic functions and the ratio $\phi_4(x+x_0)\phi_5(x_0)/\phi_5(x+$

$x_0)\phi_4(x_0)$. Then from

$$\begin{aligned}\phi_1(x+2x_0) &= \phi_4(x+x_0)\phi_5(x_0)+\phi_4(x_0)\phi_5(x+x_0)\\ &= \phi_4(x+x_0)\phi_5(x_0)\left(1+\frac{\phi_5(x+x_0)\phi_4(x_0)}{\phi_4(x+x_0)\phi_5(x_0)}\right)\\ &= e^{(\lambda_2-\lambda_1)x}\,\frac{\Phi(x;\nu_1)}{\Phi(x;\nu_2)}\phi_1(2x_0),\end{aligned}$$

we will obtain

$$\phi_4(x+x_0) = \frac{2\phi_4(x_0)e^{(\lambda_2-\lambda_1)x}}{1+\phi_5(x+x_0)\phi_4(x_0)/\phi_4(x+x_0)\phi_5(x_0)}\,\frac{\Phi(x;\nu_1)}{\Phi(x;\nu_2)}, \tag{6.2}$$

with $\phi_5(x+x_0)$ immediately following.

Now from (5.6), we obtain

$$\frac{\phi_{2k}(x+x_0)}{\phi_{2k+1}(x+x_0)}\,\frac{\phi_{2k+1}(x_0)}{\phi_{2k}(x_0)} = 1+\frac{N_k}{D_k}, \tag{6.3}$$

where

$$N_k = \frac{\phi'_{2k}(x_0)}{\phi_{2k}(x_0)}-\frac{\phi'_{2k+1}(x_0)}{\phi_{2k+1}(x_0)}, \qquad D_k = \frac{\phi'_{2k+1}(x_0)}{\phi_{2k+1}(x_0)}+\lambda_k-\frac{\Phi'(x;\nu_k)}{\Phi(x;\nu_k)}.$$

Here $N_1 = 2N_2$ and by our assumption that x_0 was a generic point, these are nonvanishing. Further, from $\phi_2(x)=\phi_4^2(x)$ and $\phi_3(x)=\phi_5^2(x)$, we see that

$$1+\frac{N_1}{D_1} = \left(1+\frac{N_2}{D_2}\right)^2.$$

Expanding this shows that $D_2^2 = (D_2+N_2/2)D_1$, which upon using (6.1) yields

$$\begin{aligned}\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2-\frac{\Phi'(x;\nu_2)}{\Phi(x;\nu_2)}\right)^2 = &\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2+\frac{N_2}{2}-\frac{\Phi'(x;\nu_2)}{\Phi(x;\nu_2)}\right)\\ &\times\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2-\frac{N_2}{2}-\frac{\Phi'(x;\nu_1)}{\Phi(x;\nu_1)}\right).\end{aligned} \tag{6.4}$$

Suppose that ν_2 is finite. Comparing the pole behavior of each side of (6.5) shows that $\nu_1=\nu_2$ and consequently that $N_2=0$, a contradiction. The remaining possibility is that ν_2 is infinite, which we now show to be a consistent solution. This can happen only if the elliptic function degenerates into a hyperbolic or trigonometric function, and without loss of generality we choose the former. In this case

$$\Phi(x;\nu) = \frac{\kappa\sinh\kappa(\nu-x)}{\sinh\kappa\nu\sinh\kappa x}e^{x\kappa\coth\kappa\nu} \quad\text{and}\quad \Phi(x;\infty)=\frac{\kappa}{\sinh\kappa x}. \tag{6.5}$$

Let us then suppose that $\nu_2=\infty$. Utilizing (A.14) and (A.15) (see the appendix), we then must solve

$$\begin{aligned}\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2+\kappa\coth\kappa x\right)^2 = &\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2+\frac{N_2}{2}+\kappa\coth\kappa x\right)\\ \times&\left(\frac{\phi'_5(x_0)}{\phi_5(x_0)}+\lambda_2-\frac{N_2}{2}+\kappa\coth\kappa(\nu_1-x)+\kappa\coth\kappa x-\kappa\coth\kappa\nu_1\right).\end{aligned}$$

This holds provided that

$$N_2^2 = \frac{4\kappa^2}{\sinh \kappa\nu_1}, \qquad \frac{\phi_5'(x_0)}{\phi_5(x_0)} + \lambda_2 + \frac{N_2}{2} + \kappa \coth\kappa\nu_1 = 0,$$

which determines ν_1, λ_2, and (via (6.1)) λ_1 in terms of $\phi_4(x_0)$, $\phi_5(x_0)$, $\phi_4'(x_0)$, and $\phi_5'(x_0)$. The choice of sign in taking the square root here is arbitrary (just defining ν_1), and we will take $N_2 = -2\kappa/\sinh \kappa\nu_1$. Substituting these into (6.3), we find that

$$\begin{aligned}\frac{\phi_4(x+x_0)}{\phi_5(x+x_0)}\frac{\phi_5(x_0)}{\phi_4(x_0)} &= \frac{\kappa \coth \kappa x - \kappa \coth \kappa\nu_1 + \frac{N_2}{2}}{\kappa \coth \kappa x - \kappa \coth \kappa\nu_1 - \frac{N_2}{2}} \\ &= \coth\left(\frac{\kappa\nu_1}{2}\right) \tanh \kappa \left(\frac{\nu_1}{2} - x\right).\end{aligned}$$

Now employing (6.5) shows

$$\phi_1(x+2x_0) = e^{(\lambda_2 - \lambda_1 + \kappa \coth \kappa\nu_1)x} \frac{\sinh \kappa(\nu_1 - x)}{\sinh \kappa\nu_1} \phi_1(2x_0), \tag{6.6}$$

where the exponential may be rewritten to yield

$$\begin{aligned}\lambda_2 - \lambda_1 + \kappa \coth \kappa\nu_1 &= \frac{\phi_4'(x_0)}{\phi_4(x_0)} - \frac{N_2}{2} + \kappa \coth \kappa\nu_1 = \frac{\phi_4'(x_0)}{\phi_4(x_0)} + \kappa \coth\left(\frac{\kappa\nu_1}{2}\right) \\ &= \frac{\phi_5'(x_0)}{\phi_5(x_0)} + \frac{N_2}{2} + \kappa \coth \kappa\nu_1 = \frac{\phi_5'(x_0)}{\phi_5(x_0)} + \kappa \tanh\left(\frac{\kappa\nu_1}{2}\right).\end{aligned}$$

We now have the information needed to determine $\phi_4(x)$ and $\phi_5(x)$ via (6.2), which gives

$$\phi_4(x+x_0) = \frac{\sinh \kappa(\frac{\nu_1}{2} - x)}{\sinh(\frac{\kappa\nu_1}{2})} e^{\left(\phi_4'(x_0)/\phi_4(x_0) + \kappa \coth(\kappa\nu_1/2)\right)x} \phi_4(x_0), \tag{6.7}$$

$$\phi_5(x+x_0) = \frac{\cosh \kappa(\frac{\nu_1}{2} - x)}{\cosh(\frac{\kappa\nu_1}{2})} e^{\left(\phi_5'(x_0)/\phi_5(x_0) + \kappa \tanh(\kappa\nu_1/2)\right)x} \phi_5(x_0). \tag{6.8}$$

Assembling this provides the following result.

LEMMA 4. *The general analytic solution to* (1.4) *is given by* (6.6), (6.7), *and* (6.8)*, where x_0 is a generic point.*

Example 3. We conclude by showing how our results determine the solutions of the functional equation (1.7):

$$\Psi_1(x+y) = \Psi_2(x+y)\phi_2(x)\phi_3(y) + \Psi_3(x+y)\phi_4(x)\phi_5(y).$$

This equation encompasses as particular cases equations (1.4) (with $\Psi_2 = \Psi_3 = 1$, $\phi_2(x) = \phi_4(x)$, and $\phi_3(x) = \phi_5(x)$) and (1.6) (with ($\phi_2(x) = \phi_3(x)$ and $\phi_4(x) = \phi_5(x)$), which have already been discussed. Because of this, we will consider only the generic case where $\phi_2(x) \neq \lambda\phi_3(x)$, $\phi_4(x) \neq \gamma\phi_5(x)$, and $\Psi_2(x) \neq \delta\Psi_3(x)$ (where λ, γ, and δ are constants) rather than these limits. Our first step is to relate (1.7) to (1.1).

LEMMA 5. *The functions $\Psi_m(x)$ $(m = 1, 2, 3)$ and $\phi_n(x)$ $(n = 2, 3, 4, 5)$ give a solution of equation* (1.7) *if and only if*

$$\frac{\Psi_3(x+y)}{\Psi_2(x+y)} = -\frac{\begin{vmatrix}\phi_2(x) & \phi_2(y) \\ \phi_3(x) & \phi_3(y)\end{vmatrix}}{\begin{vmatrix}\phi_4(x) & \phi_4(y) \\ \phi_5(x) & \phi_5(y)\end{vmatrix}} \tag{6.9}$$

and

$$\frac{\Psi_1(x+y)}{\Psi_2(x+y)} = -\frac{\begin{vmatrix} \phi_2(x)\phi_5(x) & \phi_2(y)\phi_5(y) \\ \phi_3(x)\phi_4(x) & \phi_3(y)\phi_4(y) \end{vmatrix}}{\begin{vmatrix} \phi_4(x) & \phi_4(y) \\ \phi_5(x) & \phi_5(y) \end{vmatrix}}. \tag{6.10}$$

Proof. First, assume that the functions Ψ_m and ϕ_n give a solution of equation (1.7). Then after interchanging x and y in (1.7) and subtracting the result from (1.7), we obtain equation (6.9). Upon substituting the formula for $\Psi_3(x+y)/\Psi_2(x+y)$ into (1.7), we arrive at formula (6.10).

In the other direction, let the functions Ψ_m and ϕ_n now satisfy (6.9) and (6.10). Upon writing the right-hand side of (1.7) as

$$\begin{aligned} &\Psi_2(x+y)\phi_2(x)\phi_3(y) + \Psi_3(x+y)\phi_4(x)\phi_5(y) \\ &\qquad = \Psi_2(x+y)\left(\phi_2(x)\phi_3(y) + \frac{\Psi_3(x+y)}{\Psi_2(x+y)}\phi_4(x)\phi_5(y)\right) \end{aligned} \tag{6.11}$$

and using expression (6.9) for $\Psi_3(x+y)/\Psi_2(x+y)$, we find that the term in brackets in (6.11) rearranges to give precisely the right-hand side of (6.10); substituting for this then yields (1.7) and therefore the required solution. □

We may now apply Theorem 1 to show that if the functions $\Psi_m(x)$ $(m = 1, 2, 3)$ give a solution of (1.7), then we must have the ratios

$$\frac{\Psi_1(x)}{\Psi_2(x)} = c_1 e^{\lambda_1 x}\frac{\Phi(x;\mu_1)}{\Phi(x;\mu_2)}, \qquad \frac{\Psi_3(x)}{\Psi_2(x)} = c_2 e^{\lambda_2 x}\frac{\Phi(x;\mu_3)}{\Phi(x;\mu_4)}. \tag{6.12}$$

Further, because the denominators of (6.9) and (6.10) are the same, Theorem 2 shows that $\mu_4 = \mu_2$. Theorem 1 also determines the functions $\phi_n(x)$ $(n = 2, 3, 4, 5)$ up to a $\mathcal{G}$ action. In fact, given three functions $\Psi_m(x)$ $(m = 1, 2, 3)$ whose ratios satisfy (6.12) with $\mu_4 = \mu_2$, this is also sufficient to guarantee that there are functions $\phi_n(x)$ $(n = 2, 3, 4, 5)$ for which (1.7) holds. To see this, let us substitute these ratios into equation (1.7) to give

$$\begin{aligned} c_1 e^{\lambda_1(x+y)}\Phi(x+y;\mu_1) = {}& \Phi(x+y;\mu_2)\phi_2(x)\phi_3(y) \\ & + c_2 e^{\lambda_2(x+y)}\Phi(x+y;\mu_3)\phi_4(x)\phi_5(y). \end{aligned} \tag{6.13}$$

We will have established sufficiency once we have shown how to construct the functions $\phi_n(x)$. This will be achieved by utilizing various properties of the functions $\Phi(x;\nu)$.

Lemma 6. *The Baker–Akhiezer functions $\Phi(x;\nu)$ satisfy the equations*

$$\Phi(x+\alpha;\nu) = -e^{(\zeta(\alpha-\nu)+\zeta(\nu)-\zeta(\alpha))x}\Phi(\alpha;\nu)\,\frac{\Phi(x;\nu-\alpha)}{\Phi(-x;\alpha)} \tag{6.14}$$

and

$$\begin{aligned} ce^{\gamma(x+y)}\,\Phi(x+y;\nu_1+\nu_2) = {}& \Phi(x+y;\nu_1)\,\Phi(x;\nu_2)\,\Phi(y;\nu_2) \\ & - \Phi(x+y;\nu_2)\,\Phi(x;\nu_1)\,\Phi(y;\nu_1), \end{aligned} \tag{6.15}$$

where $c = \wp(\nu_2) - \wp(\nu_1)$ and $\gamma = \zeta(\nu_1) + \zeta(\nu_2) - \zeta(\nu_1+\nu_2)$.

These follow directly from the definition of $\Phi(x;\nu)$ and the properties of the Weierstrass sigma function; in particular, (6.15) is a consequence of the "three term relation" of the sigma function [23, Chapter 20.53, Example 5].

Upon setting $x \to x+\alpha$ in (6.15), we obtain

$$\begin{aligned} ce^{\gamma(x+y+\alpha)}\Phi(x+y+\alpha;\nu_1+\nu_2) &= \Phi(x+y+\alpha;\nu_1)\Phi(x+\alpha;\nu_2)\Phi(y;\nu_2) \\ &\quad - \Phi(x+y+\alpha;\nu_2)\Phi(x+\alpha;\nu_1)\Phi(y;\nu_1). \end{aligned} \tag{6.16}$$

Now by substituting (6.14) in (6.16) and setting $\mu_1 = \nu_1+\nu_2-\alpha$, $\mu_2 = \nu_1 - \alpha$, and $\mu_3 = \nu_2 - \alpha$, after some rearrangement, we obtain

$$\begin{aligned} c'e^{\lambda'(x+y)}\Phi(x+y;\mu_1) &= \Phi(x+y;\mu_2)\frac{\Phi(x;\mu_3)}{\Phi(-x;\mu_1-\mu_2-\mu_3)}\Phi(y;\mu_1-\mu_2) \\ &\quad + c''e^{\lambda''(x+y)}\Phi(x+y;\mu_3)\frac{\Phi(x;\mu_2)}{\Phi(-x;\mu_1-\mu_2-\mu_3)}\Phi(y;\mu_1-\mu_3) \end{aligned} \tag{6.17}$$

for appropriate constants c', c'', λ', and λ''. This is precisely of the desired form (6.13). Therefore, we have shown the following.

THEOREM 3. *Given functions $\Psi_m(x)$ $(m = 1,2,3)$, there are functions $\phi_n(x)$ $(n = 2,3,4,5)$ for which the functional equation* (1.7) *is true if and only if the following ratios take place:*

$$\frac{\Psi_1(x)}{\Psi_2(x)} = c_1 e^{\lambda_1 x}\frac{\Phi(x;\mu_1)}{\Phi(x;\mu_2)}, \qquad \frac{\Psi_3(x)}{\Psi_2(x)} = c_2 e^{\lambda_2 x}\frac{\Phi(x;\mu_3)}{\Phi(x;\mu_2)}, \tag{6.18}$$

where c_m, λ_m $(m = 1,2)$, and μ_n $(n = 1,2,3)$ are free parameters.

Appendix. Elliptic functions.

A.1. The Weierstrass elliptic functions. The Weierstrass elliptic functions are based on a lattice with periods 2ω and $2\omega'$, where $\Im(\omega'/\omega) > 0$. They satisfy the homogeneity relations

$$\sigma(tx|t\omega,t\omega') = t\sigma(x|\omega,\omega'), \qquad \zeta(tx|t\omega,t\omega') = t^{-1}\zeta(x|\omega,\omega'), \tag{A.1}$$

$$\wp(tx|t\omega,t\omega') = t^{-2}\wp(x|\omega,\omega'). \tag{A.2}$$

Here $\zeta(x) = (\ln\sigma(x))'$ and $\wp(x) = -\zeta'(x)$. These homogeneity relations mean that our function $\Phi(x;\nu) \equiv \Phi(x;\nu|\omega,\omega')$ satisfies

$$\Phi(tx;t\nu|t\omega,t\omega') = t^{-1}\Phi(x;\nu|\omega,\omega').$$

The parameters g_2 and g_3, alternately used to describe the elliptic function, are given by

$$g_2 = 60\sum_{m,n\in\mathbb{Z}}{}' \,\Omega^{-4}, \qquad g_3 = 140\sum_{m,n\in\mathbb{Z}}{}' \,\Omega^{-6},$$

where $\Omega = 2m\omega + 2n\omega'$.

The Weierstrass $\wp$-function, $\wp(x) \equiv \wp(x|\omega,\omega') = \wp(x;g_2,g_3)$, satisfies the differential equation

$$\wp'^2(x) = 4\wp^3(x) - g_2\wp(x) - g_3 = 4\big(\wp^3(x) - 5c_2\wp(x) - 7c_3\big), \tag{A.3}$$

where $c_2 = g_2/20$ and $c_3 = g_3/28$. This means that

$$\wp''(x) = 6\wp^2(x) - 10c_2. \tag{A.4}$$

The terms c_k $(k \geq 4)$ in the Laurent expansion of the Weierstrass $\wp$ function,

$$\wp(x) = \frac{1}{x^2} + \sum_{l=2} c_l\, x^{2l-2}, \tag{A.5}$$

are expressible in terms of c_2 and c_3.

If $e_i \equiv e_i(\omega, \omega')$ $(i = 1, 2, 3)$ denote the roots of the cubic

$$4x^3 - g_2 x - g_3 = 0, \tag{A.6}$$

where

$$\wp'^2(x) = 4(\wp(x) - e_1)(\wp(x) - e_2)(\wp(x) - e_3),$$

the half-periods ω_i are defined by

$$\wp(\omega_i) = e_i, \quad \text{where } \omega_1 = \omega, \quad \omega_2 = \omega + \omega', \quad \text{and} \quad \omega_3 = \omega'.$$

Clearly,

$$e_i(t\omega, t\omega') = t^{-2} e_i(\omega, \omega'). \tag{A.7}$$

Assuming that g_2 and g_3 are real and that the discriminant of (A.6) is positive (the case of interest in this paper), the e_i's are real and may be ordered $e_1 \geq e_2 \geq e_3$.

A.2. The Jacobi elliptic functions. The Jacobi elliptic functions are characterized by a parameter m. Thus, for example, $\mathrm{sn}(x) \equiv \mathrm{sn}(x|m)$ has periods $4K(m)$ and $2iK'(m)$, where $K(m)$ is the complete elliptic function of the first kind. The e_i's are related to the parameter m of the Jacobi elliptic functions by

$$e_1 = \frac{2-m}{3}\frac{K^2(m)}{\omega^2}, \qquad e_2 = \frac{2m-1}{3}\frac{K^2(m)}{\omega^2}, \quad \text{and} \quad e_3 = \frac{-1-m}{3}\frac{K^2(m)}{\omega^2}. \tag{A.8}$$

Thus

$$g_2 = \frac{4}{3}(1 - m + m^2)\frac{K^4(m)}{\omega^4}, \qquad g_3 = \frac{4}{27}(m-2)(2m-1)(m+1)\frac{K^6(m)}{\omega^6}, \tag{A.9}$$

and

$$\frac{\omega}{\omega'} = \frac{iK'(m)}{K(m)}, \qquad \omega = \frac{K(m)}{\sqrt{e_1 - e_3}}. \tag{A.10}$$

We find that

$$\Phi(x;\omega') = \frac{a}{\mathrm{sn}(a\,x)}, \qquad \Phi(x;\omega) = a\,\frac{\mathrm{cn}(a\,x)}{\mathrm{sn}(a\,x)}, \quad \text{and} \quad \Phi(x;\omega+\omega') = a\,\frac{\mathrm{dn}(a\,x)}{\mathrm{sn}(a\,x)}.$$

Here $a = \sqrt{e_1 - e_3}$ converts the periods of Φ based on ω and ω' to those of the Jacobi functions based on $K(m)$ and $iK'(m)$. Equally, we may write this as

$$\Phi(x; t\omega' | t\omega, t\omega') = \frac{1}{\mathrm{sn}(x|m)}, \quad \text{with } t = \sqrt{e_1 - e_3} = \frac{K(m)}{\omega}. \tag{A.11}$$

Thus with these periods in mind, we write

$$\Phi(x; iK'(m)) = \frac{1}{\mathrm{sn}(x)}, \qquad \Phi(x; K(m) + iK'(m)) = \frac{\mathrm{dn}(x)}{\mathrm{sn}(x)}, \tag{A.12}$$

$$\Phi(x; K(m)) = \frac{\mathrm{cn}(x)}{\mathrm{sn}(x)}. \tag{A.13}$$

A.3. Degenerations. When the discriminant of (A.6) vanishes, one (or both) of the periods of the elliptic function vanishes, yielding hyperbolic, trigonometric (or rational) functions. If $e_1 = e_2 = c$, $e_3 = -2c$ (and so $g_2 = 12c^2$, $g_3 = -8c^3$), we then have

$$\sigma(x; 12c^2, -8c^3) = \frac{\sinh \kappa x}{\kappa} e^{-\kappa^2 x^2/6} \quad \text{and} \quad \wp(x; 12c^2, -8c^3) = \frac{\kappa^2}{3} + \frac{\kappa^2}{\sinh^2 \kappa x},$$

where $\kappa = \sqrt{3c}$. In this case,

$$\text{(A.14)} \quad \begin{aligned} \Phi(x;\nu) &= \frac{\kappa \sinh \kappa(\nu - x)}{\sinh \kappa\nu \sinh \kappa x} e^{x\kappa \coth \kappa\nu} = \kappa \left(\coth \kappa x - \coth \kappa\nu\right) e^{x\kappa \coth \kappa\nu}, \\ \Phi'(x;\nu) &= -\kappa\, \Phi(x;\nu) \left(\coth \kappa(\nu - x) + \coth \kappa x - \coth \kappa\nu\right). \end{aligned}$$

In particular,

$$\text{(A.15)} \qquad \Phi(x;\infty) = \frac{\kappa}{\sinh \kappa x} \quad \text{and} \quad \frac{\Phi'(x;\infty)}{\Phi(x;\infty)} = -\kappa \coth \kappa x.$$

REFERENCES

[1] H. Airault, H. McKean, and J. Moser, *Rational and elliptic solutions of the KdV equation and related many-body problems*, Comm. Pure Appl. Math., 30 (1977), pp. 95–125.

[2] H. W. Braden and V. M. Buchstaber, *Integrable systems with pairwise interactions and functional equations*, Rev. Math. and Math. Phys., to appear.

[3] M. Bruschi and F. Calogero, *The Lax representation for an integrable class of relativistic dynamical systems*, Comm. Math. Phys., 109 (1987), pp. 481–492.

[4] M. Bruschi and F. Calogero, *General analytic solution of certain functional equations of addition type*, SIAM J. Math. Anal., 21 (1990), pp. 1019–1030.

[5] V. M. Buchstaber, *Functional equations which are associated with addition theorems for elliptic function and two valued algebraic groups*, Uspekhi Mat. Nauk, 45 (1990), pp. 185–186.

[6] V. M. Buchstaber, Unpublished report, Max-Planck-Institut, Bonn, Germany, 1992.

[7] V. M. Buchstaber, G. Felder, and A. P. Veselov, *Elliptic Dunkl operators, root systems, and functional equations*, Duke Math. J., 76 (1994), pp. 885–911.

[8] V. M. Buchstaber and A. N. Khodov, *Formal groups, functional equations and generalised cohomology theories*, Mat. Sb., 181 (1990), pp. 75–94.

[9] V. M. Buchstaber and A. M. Perelomov, *On the functional equation related to the quantum three-body problem*, Contemporary Math. Physics, 175 (1996), pp. 15–34.

[10] V. M. Buchstaber and I. M. Krichever, *Vector addition theorems and Baker–Akhiezer functions*, Teor. Mat. Fiz., 94 (1993), pp. 200–212.

[11] F. Calogero, *Exactly solvable one-dimensional many-body problems*, Lett. Nuovo Cimento (2), 13 (1975), pp. 411–416.

[12] F. Calogero, *On a functional equation connected with integrable many-body problems*, Lett. Nuovo Cimento, 16 (1976), pp. 77–80.

[13] F. Calogero, *One-dimensional many-body problems with pair interactions whose ground-state wavefunction is of product type*, Lett. Nuovo Cimento, 13 (1975), pp. 507–511.

[14] F. Gliozzi and R. Tateo, *ADE functional dilogarithm identities and integrable models*, Phys. Lett., B348 (1995), pp. 84–88.

[15] F. Hirzebruch, T. Berger, and R. Jung, *Manifolds and Modular Forms*, Vieweg, Wiesbaden, Germany, 1992.

[16] I. M. Krichever, *Elliptic solutions of Kadomtsev–Petviashvili equation and integrable particle systems*, Funktsional. Anal. Prilozhen., 14 (1980), pp. 45–54.

[17] I. M. Krichever, *Generalized elliptic genera and Baker–Akhiezer functions*, Mat. Zametki., 47 (1992), p. 132.

[18] M. A. Olshanetsky and A. M. Perelomov, *Completely integrable Hamiltonian systems connected with semisimple Lie algebras*, Invent. Math., 37 (1976), pp. 93–108.

[19] S. Pydkuyko and A. Stepin, *On the solution of one functional-differential equation*, Funktsional. Anal. Prilozhen., 10 (1976), pp. 84–85 (in Russian).

[20] S. N. M. Ruijsenaars and H. Schneider, *A new class of integrable systems and its relation to solitons*, Ann. Phys. (NY), 170 (1986), pp. 370–405.

[21] S. N. M. Ruijsenaars, *Complete integrability of relativistic Calogero–Moser systems and elliptic function identities*, Comm. Math. Phys., 110 (1987), pp. 191–213.

[22] S. N. M. Ruijsenaars, *Relativistic Toda systems*, Comm. Math. Phys., 133 (1990), pp. 217–247.

[23] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1927.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 924–943, July 1997

GLOBAL ESTIMATES FOR SOLUTIONS OF PARTIAL DIFFERENTIAL EQUATIONS[*]

CHANGMEI LIU[†]

Abstract. For a polynomial $P(\xi)$ in ξ in $\mathbf{R}^n$ with constant complex coefficients, the operator defined by $R(z)f = \mathcal{F}^{-1}((P(\cdot)-z)^{-1}\hat{f})$, where $\wedge$ denotes the Fourier transform and $\mathcal{F}^{-1}$ denotes its inverse, is not bounded from L^2 to L^2 when z is in the spectrum of $P(D)$. What are suitable spaces B and C so that $R(z)$ is bounded from B to C? When $P(\xi)$ is *simply characteristic*, we prove that the operator $R(z)$ is bounded from B_s to B^*_{1-s}, $0 \leq s \leq 1$, where B_s are spaces reasonably smaller than L^2 and B^*_{1-s} are spaces reasonably larger than L^2.

Key words. solutions, estimates, simply characteristic, critical value, Besov space, Fourier transform, inverse Fourier transform

AMS subject classifications. 35E20, 35B05, 35B40

PII. S0036141095290628

1. Introduction and statements of results.

In recent years, a method for obtaining uniqueness results (and even giving a theoretical inversion procedure) for a class of inverse problems in potential scattering has been developed very successfully. Sylvester and Uhlmann treated an inverse boundary-value problem from electric impedance tomography (see [13]). Nachman and Ablowitz, Beals and Coifman, and Novikov and Henkin studied some problems in inverse scattering (see [11, 3, 12]). These have yielded a breakthrough on some problems for which only linearized approximation had been treatable before and led to the solution of a great number of related inverse problems.

Most of these works treat problems which can, by one device or another, be reduced to problems for the Schrödinger operator $-\Delta+q$. One of the crucial ingredients is a family of solutions $\phi(x,\zeta)$ of $(-\Delta+q)\phi = E\phi$ which behave like so-called inhomogeneous plane waves $\exp(ix\cdot\zeta)$ for large values of the complex vector ζ. One motivation for the study of such solutions comes from the observation that it is possible to have $\Delta e^{ix\cdot\zeta} = Ee^{ix\cdot\zeta}$ with $\zeta^2 = E$ and the energy E fixed, while ζ is made arbitrarily large. Then fixed-energy uniqueness follows from a complex version of the Born limit, first observed in [11]. Following this key idea, inverse boundary-value problems for parabolic, hyperbolic, and other more general partial differential operators were investigated by Isakov [5]. An inverse boundary-value problem for the biharmonic operator at zero energy was considered by Ikehata [6]. More related inverse boundary-value problems and inverse scattering problems can be found in [9, 7, 10] and the references therein.

The direct scattering theory for operators of the form $P_0(D) + V(x,D)$, where $P_0(D)$ is a partial differential operator with real constant coefficients and simply characteristic (see Definition 1.1 below) and V is a short-range perturbation, has been developed successfully by Agmon and Hörmander (see [4]) and many other authors. One of the key ingredients of such a theory is an estimate for the resolvent $(P_0(D) - z)^{-1}$ which remains valid as z approaches the real axis. This clearly cannot happen on

[*] Received by the editors August 15, 1995; accepted for publication (in revised form) February 28, 1996.

http://www.siam.org/journals/sima/28-4/29062.html

[†] Biometrics Unit, Cornell University, 432 Warren Hall, Ithaca, NY 14853 (cl103@cornell.edu).

the space L^2 when z approaches the spectrum of $P_0(D)$, but if $R_0(z) = (P_0(D)-z)^{-1}$ is viewed as an operator from a suitable space X smaller than L^2 to another space Y larger than L^2, then its norm can be shown to remain bounded, independent of the distance from z to the real axis (if z stays in a bounded subset K contained in the complex plane $\mathbf{C}$),

$$\|(P_0(D)-z)^{-1}f\|_Y \leq C\|f\|_X. \tag{1.1}$$

The result of the form of (1.1) is known as the "limiting-absorption principle." To see what spaces X and Y are appropriate, we note, for example, that when $P_0(D) = -\Delta$ in $\mathbf{R}^3$, the operator $(P_0(D)-z^2)^{-1}$, where $z \in \mathbf{C}\backslash\{0\}$, corresponds to convolution by

$$G_0^+(x-y) = \frac{e^{iz|x-y|}}{4\pi|x-y|}.$$

Even for $f \in C_0^\infty$,

$$G_0^+ * f = O(1/|x|) \quad \text{as } |x| \to \infty.$$

Thus $G_0^+ * f$ is in general not in L^2, but it can be shown that $(1+x^2)^{-\frac{\delta}{2}}(G_0^+ * f)$ does belong to L^2 whenever $(1+x^2)^{\frac{\delta}{2}} f$ is in L^2 if $\delta > 1/2$. This motivates the introduction of the weighted L^2 spaces:

$$L_\delta^2 = \left\{ v \in L^2(\mathbf{R}^n) : \int_{\mathbf{R}^n} (1+|x|^2)^\delta |v(x)|^2 dx < \infty \right\}.$$

Estimates of the form (1.1), with $X = L_\delta^2$, $Y = L_{-\delta}^2$, $\delta > 1/2$, and P_0 simply characteristic, were first proved by Agmon in [1]. In [2], Agmon and Hörmander showed that the following class of spaces B_s (the Fourier transform of Besov space $B_2^{s,1}$) and their duals B_s^* (the Fourier transform of Besov space $B_2^{-s,\infty}$) $(-\infty < s < \infty)$,

$$B_s = \left\{ v \in L_{\text{loc}}^2(\mathbf{R}^n) : \sum_{j=1}^\infty R_j^s \left(\int_{\Omega_j} |v|^2 dx \right)^{1/2} < \infty \right\}, \tag{1.2}$$

$$B_s^* = \left\{ u \in L_{\text{loc}}^2(\mathbf{R}^n) : \sup_{j\geq 1} R_j^{-s} \left(\int_{\Omega_j} |u|^2 dx \right)^{1/2} < \infty \right\}, \tag{1.3}$$

where

$$R_0 = 0, \qquad R_j = 2^{j-1}, \quad j = 1, 2, \dots,$$

$$\Omega_j = \{x \in \mathbf{R}^n : R_{j-1} < |x| < R_j\}, \quad j = 1, 2, \dots,$$

capture quite precisely the behavior of the resolvent operator at infinity. The relationship between L_δ^2 and B_s is

$$L_\delta^2 \subset B_s \subset L_s^2 \quad \text{and} \quad L_{-s}^2 \subset B_s^* \subset L_{-\delta}^2$$

for $\delta > s \geq 0$.

Definition 1.1. *Let $P(\xi)$ be a real-valued polynomial of degree m in $\xi \in \mathbf{R}^n$ such that*

$$\Lambda(P_0) = \{\eta \in \mathbf{R}^n : P_0(\xi+\eta) \equiv P_0(\xi)\} = \{0\}.$$

P_0 will be called simply characteristic if

$$\tilde{P}_0(\xi) \le C \left(\sum_{|\alpha|\le 1} |P_0^{(\alpha)}(\xi)| + 1 \right), \quad \xi \in \mathbf{R}^n, \tag{1.4}$$

where

$$\tilde{P}_0(\xi) = \sum_{0\le|\alpha|\le m} |P_0^{(\alpha)}(\xi)|.$$

To study fixed-energy inverse problems in potential scattering for a general class of differential operators $P_0(D)$, we want solutions $\phi(x,\zeta)$ of $(P_0(D)+q)\phi = \lambda\phi$ which behave like $e^{ix\cdot\zeta}$ with $\zeta \in \mathbf{C}^n$, $P_0(\zeta) = \lambda$. The construction of such solutions requires a generalized limiting-absorption estimate for $P_0(D,\zeta) = P_0(D+\zeta) - \lambda$. The first such estimate was obtained by Sylvester and Uhlmann [13] for the Laplacian $-\Delta$ at zero energy ($\zeta^2 = 0$):

$$\|(-\Delta - 2i\zeta\cdot\nabla)^{-1} f\|_{L^2_{-\delta}} \le \frac{C}{|\zeta|} \|f\|_{L^2_{1-\delta}}, \quad 0<\delta<1. \tag{1.5}$$

To obtain an estimate of the form of (1.5) for a general class of differential operators $P_0(D,\zeta) = P_0(D+\zeta) - \lambda$, we first need an analogue of the Agmon–Hörmander estimate (1.1) for complex polynomials. Our main goal in this paper is to give such an estimate. Note that the key point for the validity of estimate (1.5) is that the real and imaginary parts of the symbol $(\xi^2 + 2\xi\cdot\zeta)$ of the differential operator $(-\Delta - 2i\zeta\cdot\nabla)$ with $\zeta = 0$ are linearly independent on the zero set $\{\xi \in \mathbf{R}^n, \xi^2 + 2\xi\cdot\zeta = 0\}$, which allows one to reduce the resolvent operator $(-\Delta - 2i\zeta\cdot\nabla)^{-1}$ to the inverse of the Cauchy–Riemann operator. Inspired by the same spirit, we first classify a class of *simply characteristic* polynomials (see Definition 1.3 below) and then obtain an estimate similar to (1.1) for such complex polynomials. For clarity, let us stress that emphasized *simply characteristic* refers to Definition 1.3 and nonemphasized simply characteristic refers to Definition 1.1 given by Agmon and Hörmander. We also distinguish *critical values* in Definition 1.4 from critical values in the normal sense. We will denote by $\mathcal{F}$ or $\hat{\ }$ the Fourier transform and by $\mathcal{F}^{-1}$ or $\vee$ its inverse. Our main estimate in this paper is as follows.

Theorem 1.2. *Assume that P is* simply characteristic *and let K be a compact subset of $\mathbf{C}$ containing no* critical value *of P in the sense given in Definition* 1.4 *below. If $f \in B_s$, $0 \le s \le 1$, it follows that $R(z)f = \mathcal{F}^{-1}((P(\cdot)-z)^{-1}\hat{f})$ belongs to B^*_{1-s} for $z \in K$ and we have the bound*

$$\|R(z)f\|_{B^*_{1-s}} \le C(n,s,c_P) \sup_{\xi\in\mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi)} \|f\|_{B_s}, \quad z \in K, \tag{1.6}$$

where $C(n,s,c_P)$ depends only on the dimension n, s, and the constant c_P in condition (1.7) *in Definition* 1.3.

Definition 1.3. *Let $P(\xi) = P_1(\xi) + iP_2(\xi)$, $\xi \in \mathbf{R}^n$, be an mth-order polynomial with complex coefficients. We define a* simply characteristic *polynomial P to be one that satisfies*

$$\tilde{P}_*(\xi) \leq c_P(|P(\xi) - z_0| + |\nabla P(\xi)|_*) \tag{1.7}$$

for all $\xi \in \mathbf{R}^n$ and some $z_0 \in \mathbf{C}$, where

$$\tilde{P}_*(\xi) = \sum_{|\alpha| \leq m, |\alpha| \neq 1} |P^{(\alpha)}(\xi)| + |\nabla P(\xi)|_*$$

and

$$|\nabla P(\xi)|_* \overset{\text{def.}}{=} \left[\sum_{i \neq j, 1 \leq i,j \leq m} \left| \det \begin{pmatrix} \frac{\partial P_1}{\partial \xi_i} & \frac{\partial P_1}{\partial \xi_j} \\ \frac{\partial P_2}{\partial \xi_i} & \frac{\partial P_2}{\partial \xi_j} \end{pmatrix} \right|^2 \right]^{1/4}. \tag{1.8}$$

Definition 1.4. *If $\nabla P_1(\xi)$ and $\nabla P_2(\xi)$ are not linearly independent at some point $\xi \in \{\xi \in \mathbf{R}^n : P(\xi) - z = 0\}$, we say that the value z is a* critical value *of P.*

Note that the estimates in Theorem 1.2 are from spaces B_s to spaces B^*_{1-s} for all $0 \leq s \leq 1$. In particular, two endpoints $s = 0$ and $s = 1$ are included, i.e., $(P(D) - z)^{-1}$ is bounded from B_1 to B^*_0 and from B_0 to B^*_1. As in direct-scattering theory, the spaces B_s and B^*_{1-s} capture the behavior of the operator $R(z)f$ at infinity more precisely than weighted L^2 spaces do. Estimates in (1.6) from B_s to B^*_{1-s} for $R(z)f$ also match the "B_s-version limiting-absorption principle" better than weighted L^2 estimates do.

The idea of the proof of Theorem 1.2 basically follows the same idea given by Sylvester and Uhlmann for the operator $(-\Delta - 2i\zeta \cdot \nabla)^{-1}$ in [13]. Using a partition of unity and a change of variables, we reduce the problem to the Cauchy–Riemann equation, for which the symbol is $\xi_1 + i\xi_2$. However, since we deal with estimates in the spaces B_s and B^*_{1-s} instead of weighted L^2 spaces—in particular, the endpoint spaces B_1 and B^*_0—the techniques we use are quite different from those that Sylvester and Uhlmann used. To generalize the estimates to the class of *simply characteristic* polynomials, we follow the approach used by Agmon and Hörmander in obtaining the B_s-version limiting-absorption principle [2].

The details of the proof of Theorem 1.2 are given in sections 2–4. In section 5, we briefly discuss the behavior of the resolvent of a *simply characteristic* differential operator $P(D)$ at infinity, which is a simple application of estimate (1.6).

Estimates of the form of (1.5), the construction of exponentially growing solutions for $P_0(D, \zeta) = P_0(D + \zeta) - \lambda$, and a uniqueness result for a general class of inverse problems were also studied by the author. In particular, an estimate similar to (1.5) for the Laplacian operator Δ at fixed nonzero energy is an immediate consequence of such estimates for a general class of partial differential operators. Then combining the estimate for Δ with estimate (1.6) will derive an estimate like (1.5) for the biharmonic operator Δ^2 at fixed nonzero energy. Therefore, exponentially growing solutions for Δ^2 can be constructed as well. We refer to [8] for the technical details and more applications.

2. Estimates for the model $1/(x_1 + ix_2)$. In this section, we will obtain the main estimate for the special model $1/(x_1 + ix_2)$. First, we recall that the norm for

$v \in B_s$ (introduced in (1.2)) is

$$\|v\|_{B_s} = \sum_{j=1}^{\infty} R_j^s \left(\int_{\Omega_j} |v|^2 dx \right)^{1/2}$$

and the norm for $u \in B_s^*$ (see (1.3)) is defined as

$$\|u\|_{B_s^*} = \sup_{j \geq 1} R_j^{-s} \left(\int_{\Omega_j} |u|^2 dx \right)^{1/2}.$$

For $s > 0$, since

$$\|u\|_{B_s^*}^2 \leq \sup_{R \geq 1} R^{-2s} \int_{|x|<R} |u|^2 dx \leq \frac{2^{2s}}{(1-2^{-2s})} \|u\|_{B_s^*}^2,$$

the norm $\|u\|_{B_s^*}$ is equivalent to $[\sup_{R\geq 1} R^{-2s} \int_{|x|<R} |u|^2 dx]^{1/2}$.

THEOREM 2.1. *Let $f \in B_1(\mathbf{R}^n)$, $n \geq 2$. Define*

$$u = \left(\frac{1}{\xi_1 + i\xi_2} \hat{f}(\xi) \right)^{\vee} = \frac{i}{2\pi} f * \left(\frac{1}{x_1 + ix_2} \right), \tag{2.1}$$

where $*$ represents the convolution with respect to the first two variables. Then there is a constant $C > 0$, depending only on the dimension n, such that

$$\|u\|_{B_0^*} \leq C \|f\|_{B_1}. \tag{2.2}$$

Proof. Write $x = (x', x'')$, where $x' = (x_1, x_2)$ and $x'' = (x_3, \ldots, x_n)$. Then

$$u(x) = \frac{i}{2\pi} \int_{y_1, y_2} \frac{f(y, x'')}{(x_1 - y_1) + i(x_2 - y_2)} dy.$$

To prove (2.2), we first prove the following two lemmas.

LEMMA 2.2. *Assume that $f \in L^1(\mathbf{R}^2) \cap L^2_{\text{loc}}(\mathbf{R}^2)$. Let u be defined by (2.1). Then*

(i)

$$\int_{|x|<R_j} |u(x)|^2 dx \leq C \left[R_j^2 \int_{|y| \leq R_{j+1}} |f(y)|^2 dy + \|f\|_{L^1}^2 \right],$$

and for any integer m with $j - 1 \geq m \geq 0$, we have

(ii)

$$\int_{R_{j-m-1}<|x_1|<R_j} \int_{|x_2|<R_j} |u(x)|^2 dx \leq C \left[R_j^2 \int_{\Omega_j^*} |f(x)|^2 dx + (4^{m+2}) \|f\|_{L^1}^2 \right],$$

where $\Omega_j^* = \{(x_1, x_2) \in \mathbf{R}^2 : R_{j-m-2} \leq |x_1| \leq R_{j+1}, |x_2| \leq R_{j+1}\}$ and the numbers C in both estimates are constants independent of j and m.

Proof. To prove estimate (i), we note that

$$\begin{aligned}\int_{|x|<R_j} |u(x)|^2 dx &= \frac{1}{4\pi^2} \int_{|x|<R_j} \Big[\int_{|y|\le R_{j+1}} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \\ &\qquad + \int_{|y|>R_{j+1}} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big]^2 dx \\ &\le \frac{1}{2\pi^2} \Big[\int_{|x|<R_j} \Big| \int_{|y|\le R_{j+1}} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big|^2 dx \\ &\qquad + \int_{|x|<R_j} \Big| \int_{|y|>R_{j+1}} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big|^2 dx \Big].\end{aligned}$$

For the first integral, we have

$$\begin{aligned}&\int_{|x|<R_j} \Big| \int_{|y|\le R_{j+1}} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big|^2 dx \\ &\quad = \underbrace{\int_{|x|<R_j} \Big| \int_{|y|\le R_{j+1}}}_{|x-y|\le 3R_j} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big|^2 dx \\ &\quad \text{(by Young's inequality)} \\ &\quad \le \left(\int_{|t|<3R_j} \frac{1}{|t|} dt \right)^2 \int_{|y|\le R_{j+1}} |f(y)|^2 dy \\ &\quad \le (6\pi)^2 R_j^2 \int_{|y|\le R_{j+1}} |f(y)|^2 dy.\end{aligned}$$

For the second integral, we have

$$\begin{aligned}&\underbrace{\int_{|x|<R_j} \Big| \int_{|y|>R_{j+1}}}_{|x-y|\ge R_j} \frac{f(y)}{(x_1-y_1)+i(x_2-y_2)} dy \Big|^2 dx \\ &\qquad \le \frac{1}{R_j^2} \int_{|x|<R_j} \left(\int_{\mathbf{R}^2} |f(y)| dy \right)^2 dx \\ &\qquad = \pi \left(\int_{\mathbf{R}^2} |f(y)| dy \right)^2.\end{aligned}$$

Combining the two terms, we have

$$\begin{aligned}\int_{|x|<R_j} |u(x)|^2 dx &\le \frac{1}{2\pi^2} \left[(6\pi)^2 R_j^2 \int_{|y|\le R_{j+1}} |f(y)|^2 dy + \pi \|f(\cdot)\|^2_{L^1(\mathbf{R}^2)} \right] \\ &\le 18 R_j^2 \int_{|y|\le R_{j+1}} |f(y)|^2 dy + \frac{1}{2\pi} \|f(\cdot)\|^2_{L^1(\mathbf{R}^2)}.\end{aligned}$$

To prove the second estimate, we rewrite the integral in the left-hand side as:

$$\int_{R_{j-m-1}\le|x_1|\le R_j} \int_{|x_2|\le R_j} |u(x)|^2 dx$$

$$= \frac{1}{4\pi^2}\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}\left|\int_{\mathbf{R}^2}\frac{f(y)}{(x_1-y_1)+i(x_2-y_2)}dy\right|^2 dx$$
$$= \frac{1}{4\pi^2}\left[\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}\left|\left(\int_{\Omega_j^*}+\int_{\mathbf{R}^2\backslash\Omega_j^*}\right)\frac{f(y)}{(x_1-y_1)+i(x_2-y_2)}dy\right|^2 dx\right].$$

For the first integral, we have by Young's inequality that

$$\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}\left|\int_{\Omega_j^*}\frac{f(y)}{(x_1-y_1)+i(x_2-y_2)}dy\right|^2 dx$$
$$\le\left(\int_{|t|\le 6R_j}\frac{1}{|t|}dt\right)^2\int_{\Omega_j^*}|f(y)|^2dy$$

since when $R_{j-m-1}<|x_1|<R_j$, $|x_2|\le R_j$, and $y\in\Omega_j^*$,

$$|x-y|\le|x_1-y_1|+|x_2-y_2|\le|x_1|+|y_1|+|x_2|+|y_2|=6R_j.$$

For the second integral, we have

$$\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}\left|\int_{\mathbf{R}^2\backslash\Omega_j^*}\frac{f(y)}{(x_1-y_1)+i(x_2-y_2)}dy\right|^2 dx$$
$$\le\frac{1}{R_{j-m-2}^2}\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}\left(\int_{\mathbf{R}^2\backslash\Omega_j^*}|f(y)|dy\right)^2dx$$
$$\le\frac{R_j^2}{R_{j-m-2}^2}\|f(\cdot)\|^2_{L^1(\mathbf{R}^2)},$$

since when $R_{j-m-1}<|x_1|<R_j$, $|x_2|\le R_j$, and $y\in\mathbf{R}^2\backslash\Omega_j^*$,

$$|x-y|\ge|x_1-y_1|\ge R_{j-m-2}.$$

Combing the two terms, we have

$$\int_{R_{j-m-1}\le|x_1|\le R_j}\int_{|x_2|\le R_j}|u(x)|^2dx$$
$$\le\left(\frac{144}{2}\right)R_j^2\int_{\Omega_j^*}|f(y)|^2dy+\frac{1}{2\pi^2}\frac{R_j^2}{R_{j-m-2}^2}\|f(\cdot)\|^2_{L^1(\mathbf{R}^2)}$$
$$=72R_j^2\int_{\Omega_j^*}|f(x)|^2dx+\frac{1}{2\pi^2}(4^{m+2})\|f(\cdot)\|^2_{L^1(\mathbf{R}^2)}.\qquad\square$$

LEMMA 2.3. *Let $n\ge 3$ and let $x=(x',x'')$, where $x'=(x_1,x_2)$ and $x''=(x_3,\dots,x_n)$. If $f\in B_1(\mathbf{R}^n)$, then*

(i) $\int_{\mathbf{R}^2}\|f(x',\cdot)\|_{L^2(\mathbf{R}^{n-2})}dx'\le\sqrt{\pi}\|f\|_{B_1}$ *and*

(ii) $(\int_{\mathbf{R}^{n-2}}\|f(\cdot,x'')\|^2_{L^1(\mathbf{R}^2)}dx'')^{1/2}\le\sqrt{\pi}\|f\|_{B_1}$.

Proof. (i) Let $f_j=f$ in Ω_j and $f_j=0$ elsewhere. Then by the Cauchy–Schwartz inequality,

$$\int_{\mathbf{R}^2}\|f_j(x',\cdot)\|_{L^2(\mathbf{R}^{n-2})}dx'\le(\pi R_j^2)^{1/2}\left(\int_{\mathbf{R}^2}\|f_j(x',\cdot)\|^2_{L^2(\mathbf{R}^{n-2})}dx'\right)^{1/2}$$
$$\le\sqrt{\pi}R_j\left(\int_{\Omega_j}|f(x)|^2dx\right)^{1/2}.$$

Thus since $f = \sum_j f_j$, we obtain

$$\int_{\mathbf{R}^2} \|f(x',\cdot)\|_{L^2(\mathbf{R}^{n-2})} dx' \le \sum_j \int_{\mathbf{R}^2} \|f_j(x',\cdot)\|_{L^2(\mathbf{R}^{n-2})} dx' \le \sqrt{\pi}\|f\|_{B_1},$$

where the second inequality follows from the definition of the B_1 norm.

(ii) The Minkowski's inequality for integrals shows that

$$\left(\int_{\mathbf{R}^{n-2}} \|f(\cdot,x'')\|^2_{L^1(\mathbf{R}^2)} dx''\right)^{1/2} \le \int_{\mathbf{R}^2} \|f(x',\cdot)\|_{L^2(\mathbf{R}^{n-2})} dx',$$

and then (ii) follows from (i). □

Lemma 2.3 shows that the norm in B_1 is a majorant for the above mixed L^1 and L^2 norms (compare with Theorem 14.1.2 in [4]).

Now we are in a position to prove (2.2). By the definition of the B_0^* norm, we need to show that $\sup_j \int_{\Omega_j} |u|^2 dx$ can be bounded by $C\|f\|_{B_1}$. For dimension $n = 2$, (2.2) follows immediately from Lemma 2.2(i). For $n \ge 3$, we cover Ω_j in the following way:

$$\begin{aligned} \Omega_j &= \{x \in \mathbf{R}^n :\ R_{j-1} < |x| < R_j\} \\ &\subset \cup_{k=1}^n \{x \in \mathbf{R}^n : a_j < |x_k| < R_j,\ |x_l| < R_j,\ 1 \le l \le n,\ l \ne k\} \\ &\overset{\text{def.}}{=} \sum_{k=1}^n W_k^j, \end{aligned} \tag{2.3}$$

where $a_j = R_{j-1}/\sqrt{n}$ and for each k,

$$W_k^j \subset \{x \in \mathbf{R}^n : a_j < |x| < b_j\} \tag{2.4}$$

with $b_j = \sqrt{n}R_j$ (see Figure 2.1).

Let m be a positive integer such that $2^{m-1} \le \sqrt{n} \le 2^m$. Then

$$\begin{aligned} \int_{\Omega_j} |u(x',x'')|^2 dx &\le \sum_{k=1}^n \int_{W_k^j} |u(x',x'')|^2 dx \\ &= \sum_{k=1}^n \int_{a_j<|x_k|<R_j} \int_{|x_l|<R_j, l\ne k} |u(x',x'')|^2 dx_1 \cdots d\hat{x}_k \cdots dx_n dx_k. \end{aligned} \tag{2.5}$$

If $k = 1$ or 2—say $k = 1$—then

$$\begin{aligned} &\int_{a_j<|x_1|<R_j} \int_{|x_l|<R_j, l\ge 2} |u(x',x'')|^2 dx_1 \cdots dx_n \\ &\qquad \le C_1 \int_{|x''|<R_j} \left(\int_{R_{j-m-1}<|x_1|<R_j} \int_{|x_2|<R_j} |u(x',x'')|^2 dx'\right) dx'' \\ &\qquad\qquad \text{(by Lemma 2.2(ii))} \\ &\qquad \le C_2 \int_{|x''|<R_j} \left[R_j^2 \int_{\Omega_j^*} |f(y,x'')|^2 dy + \|f(\cdot,x'')\|^2_{L^1(\mathbf{R}^2)}\right] dx'' \\ &\qquad\qquad \text{(by(2.4))} \\ &\qquad \le C_2 R_j^2 \int_{R_{j-m-2}<|(y,x'')|<\sqrt{n}R_{j+1}} |f(y,x'')|^2 dy dx'' \end{aligned} \tag{2.6}$$

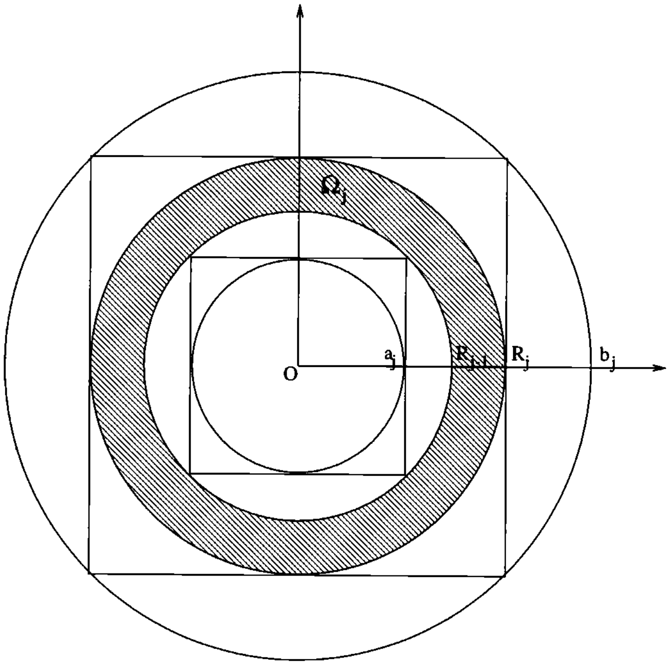


FIG. 2.1. *The graph of* Ω_j.

$$
\begin{aligned}
&+C_2 \int_{|x''|<R_j} \|f(\cdot, x'')\|^2_{L^1(\mathbf{R}^2)} dx'' \\
&\text{(by Lemma 2.3(ii))} \\
\le\ & C(4^{m+2})\|f\|^2_{B_1(\mathbf{R}^n)} + C\|f\|^2_{B_1(\mathbf{R}^n)}.
\end{aligned}
$$

If $k \geq 3$—say $k = 3$—then

$$
\begin{aligned}
&\int_{R_{j-1}/\sqrt{n}<|x_3|<R_j} \int_{|x_l|<R_j, l\neq 3} |u(x', x'')|^2 dx_1 \cdots d\hat{x}_3 \cdots dx_n dx_3 \\
&\quad \le C_3 \int_{R_{j-m-1}<|x_3|<R_j} \int_{|x_l|<R_j, l\neq 3,1,2} \left(\int_{|x'|<R_j} |u(x', x'')|^2 dx' \right) dx'' \\
&\quad \text{(by Lemma 2.2(i))} \\
&\quad \le C_3 \int_{R_{j-m-1}<|x_3|<R_j} \int_{|x_l|<R_j, l\neq 3,1,2} \left[18R_j^2 \int_{|y|<R_{j+1}} |f(y, x'')|^2 dy \right. \\
&\quad \left. + \frac{1}{2\pi} \|f(\cdot, x'')\|^2_{L^1(\mathbf{R}^2)} \right] dx'' \\
&\quad \text{(by (2.4))} \qquad (2.7) \\
&\quad \le C_3 R_j^2 \int_{R_{j-m-1}<|(y,x'')|<\sqrt{n}R_{j+1}} |f(y, x'')|^2 dy dx'' \\
&\quad + C_3 \int_{|x''|<R_j} \|f(\cdot, x'')\|^2_{L^1(\mathbf{R}^2)} dx'' \\
&\quad \text{(by Lemma 2.3(ii))} \\
&\quad \le C_3(4^{m+2})\|f\|^2_{B_1(\mathbf{R}^n)} + C_3\|f\|^2_{B_1(\mathbf{R}^n)}.
\end{aligned}
$$

Thus for each k,

$$\int_{W_k^j} |u(x',x'')|^2 dx \leq C_k(4^{m+2}\|f\|^2_{B_1(\mathbf{R}^n)} + \|f\|^2_{B_1(\mathbf{R}^n)}), \tag{2.8}$$

with m depending only on the dimension n.

Returning to (2.5), we obtain

$$\|u\|_{B_0^*} = \sup_{j\geq 1}\left(\int_{\Omega_j} |u(x)|^2 dx\right)^{1/2} \leq C\|f\|_{B_1(\mathbf{R}^n)}. \qquad \square$$

COROLLARY 2.4. *If $f \in B_0$ and $u = f * (1/(x_1 + ix_2))$, then*

$$\|u\|_{B_1^*} \leq C\|f\|_{B_0}.$$

Proof. Note that B_1^* is the dual space of B_1. Thus

$$\begin{aligned}
\|u\|_{B_1^*} &= \sup_{\|g\|_{B_1}=1}\left|\int u(x)\overline{g(x)}dx\right| \\
&= \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\left|\int_{x''}\int_{x'}\left(\int_y \frac{f(y,x'')}{(x_1-y_1)+i(x_2-y_2)}dy\right)\overline{g(x',x'')}dx'dx''\right| \\
&= \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\left|\int_{x''}\int_{x'}\int_y \frac{f(y,x'')\overline{g(x',x'')}}{(x_1-y_1)+i(x_2-y_2)}dydx'dx''\right| \\
&\quad \text{(by Fubini's Theorem)} \\
&= \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\left|\int_{x''}\int_y\left(\int_{x'} \frac{\overline{g(x',x'')}}{(x_1-y_1)+i(x_2-y_2)}dx'\right)f(y,x'')dydx''\right| \\
&\leq \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\sum_{j=1}^{\infty}\int_{\Omega_j}|f(y,x'')|\left|\int_{x'} \frac{\overline{g(x',x'')}}{(x_1-y_1)+i(x_2-y_2)}dx'\right|dydx'' \\
&\leq \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\sum_{j=1}^{\infty}\left(\int_{\Omega_j}|f(y,x'')|^2dydx''\right)^{1/2} \\
&\times\left(\int_{\Omega_j}\left|\int_{x'} \frac{\overline{g(x',x'')}}{(x_1-y_1)+i(x_2-y_2)}dx'\right|^2dydx''\right)^{1/2} \\
&\leq \frac{1}{2\pi}\sup_{\|g\|_{B_1}=1}\left[\sum_{j=1}^{\infty}\left(\int_{\Omega_j}|f(y,x'')|^2dydx''\right)^{1/2}\right. \\
&\times\left.\sup_{j\geq 1}\left(\int_{\Omega_j}\left|\int_{x'} \frac{\overline{g(x',x'')}}{(x_1-y_1)+i(x_2-y_2)}dx'\right|^2dydx''\right)^{1/2}\right] \\
&\quad \text{(by Theorem 2.1)} \\
&\leq \left[\sum_{j=1}^{\infty}\int_{\Omega_j}|f(y,x'')|^2dydx''\right)^{1/2}\Bigg]\sup_{\|g\|_{B_1}=1}(C\|\bar{g}\|_{B_1}) \\
&= C\|f\|_{B_0}.
\end{aligned}$$

Thus $u = f * (1/(x_1 + ix_2))$ is in B_1^*. $\square$

Corollary 2.5. *Let $0 < s < 1$. If $f \in B_{1-s}$, then $f * (1/(x_1 + ix_2)) \in B_s^*$ and*

$$\left\| f * \left(\frac{1}{x_1 + ix_2} \right) \right\|_{B_s^*} \leq C_s \|f\|_{B_{1-s}},$$

where C_s depends only on s and the dimension n.

Proof. First if $u \in B_0^*$, we have

$$\begin{aligned}
\|u\|_{L^2_{-s}}^2 &= \int (1 + |x|^2)^{-s} |u(x)|^2 dx \\
&= \sum_{j=1}^{\infty} \int_{\Omega_j} (1 + |x|^2)^{-s} |u(x)|^2 dx \\
&\leq \sum_{j=1}^{\infty} R_{j-1}^{-2s} \int_{\Omega_j} |u(x)|^2 dx \\
&= 2^{2s} \sum_{j=1}^{\infty} R_j^{-2s} \int_{\Omega_j} |u(x)|^2 dx \\
&\leq 4 \left[\sum_{j \leq k} R_j^{-2s} \int_{\Omega_j} |u(x)|^2 dx + \sum_{j > k} R_j^{-2s} \int_{\Omega_j} |u(x)|^2 dx \right] \\
&= 4 \left[\sum_{j \leq k} R_j^{2-2s} \frac{1}{R_j^2} \int_{\Omega_j} |u(x)|^2 dx + \sum_{j > k} R_j^{-2s} \int_{\Omega_j} |u(x)|^2 dx \right] \\
&\leq C \left[R_k^{2-2s} \frac{1}{1 - 2^{-(2-2s)}} \|u\|_{B_1^*}^2 + R_k^{-2s} \|u\|_{B_0^*}^2 \right]
\end{aligned}$$

for each $k \geq 1$, where C is independent of k. Thus

$$\|u\|_{L^2_{-s}} \leq \sqrt{C_s}/2 [R_k^{1-s} \|u\|_{B_1^*} + R_k^{-s} \|u\|_{B_0^*}].$$

Now let $f = \sum_{k=1}^{\infty} f_k$, where $f_k = f|_{\Omega_k}$. Then with $u_k = f_k * 1/(x_1 + ix_2)$,

$$\begin{aligned}
\|u\|_{L^2_{-s}} &= \left\| \sum_{k=1}^{\infty} u_k \right\|_{L^2_{-s}} \\
&\leq \sum_{k=1}^{\infty} \|u_k\|_{L^2_{-s}} \\
&\leq C_s \sum_k [R_k^{1-s} \|u_k\|_{B_1^*} + R_k^{-s} \|u_k\|_{B_0^*}] \\
&\quad \text{(by Theorem 2.1 and Corollary 2.4)} \\
&\leq C_s \sum_k [R_k^{1-s} \|f_k\|_{B_0} + R_k^{-s} \|f_k\|_{B_1}] \\
&= C_s \sum_k \left[R_k^{1-s} \left(\int_{\Omega_k} |f(x)|^2 dx \right)^{1/2} + R_k^{-s} R_k \left(\int_{\Omega_k} |f(x)|^2 dx \right)^{1/2} \right] \\
&= C_s \sum_k R_k^{1-s} \left(\int_{\Omega_k} |f(x)|^2 dx \right)^{1/2} \\
&= C_s \|f\|_{B_{1-s}}.
\end{aligned}$$

Since $L^2_{-s} \subset B^*_s$,

$$\|u\|_{B^*_s} \le C\|u\|_{L^2_{-s}} \le C_s\|f\|_{B_{1-s}}. \qquad \square$$

Note that although the map $f \to f * (1/(x_1 + ix_2))$ is bounded from L^2_δ to $L^2_{-1+\delta}$ for $0 < \delta < 1$, it is bounded from neither L^2_1 to L^2_0 nor L^2_0 to L^2_{-1}. Thus the spaces B_s provide appropriate endpoint substitutes for the weighted L^2 spaces.

3. Estimate for general distribution $(1/(H_1(\xi) + iH_2(\xi)))$.

We now generalize the result for the particular model $1/(x_1 + ix_2)$ to a general distribution $(1/(H_1(\xi)+ iH_2(\xi)))^\vee(x)$. Before doing so, we need some additional definitions and lemmas, which can be found in [4, Chapter 14].

Let $c_1, c_2, \ldots$ be a sequence of positive numbers such that for some constant $M > 0$,

$$\frac{c_j}{M} \le c_{j+1} \le Mc_j, \quad j = 1, 2, \ldots. \tag{3.1}$$

Define

$$B_{\{c\}} = \left\{ v \in L^2_{\text{loc}}(\mathbf{R}^n) : \sum_{j=1}^\infty c_j \left(\int_{\Omega_j} |v|^2 dx \right)^{1/2} < \infty \right\}.$$

Then its dual space is

$$B^*_{\{c\}} = \left\{ u \in L^2_{\text{loc}}(\mathbf{R}^n) : \sup_{j \ge 1} c_j^{-1} \left(\int_{\Omega_j} |u|^2 dx \right)^{1/2} < \infty \right\}.$$

Lemma 3.1. *Let N be the smallest integer such that $2^N > M$. Then there is a constant C_M such that if*

$$T : \ L^2_{-N} \to L^2_{-N}$$

is bounded and

$$T : \ L^2_N \to L^2_N$$

is bounded with both norms $\le A$, it follows that

$$T : \ B_{\{c\}} \to B_{\{c\}}$$

is bounded with norm $\le C_M A$.

Lemma 3.2. *Let $r \in C^N(\mathbf{R}^n)$ and assume that $D^\alpha r$ is bounded when $|\alpha| \le N$. Then the operator $r(D) = \mathcal{F}^{-1} r \mathcal{F}$ is bounded in $B_{\{c\}}$ and*

$$\|r(D)u\|_{B_{\{c\}}} \le C_M \sum_{|\alpha| \le N} \sup |D^\alpha r| \|u\|_{B_{\{c\}}}, \quad u \in B_{\{c\}},$$

where $\mathcal{F}$ is the Fourier transform operator.

Lemma 3.3. *Let X_1 and X_2 be open sets in $\mathbf{R}^n$ and Ψ be a C^{N+1} diffeomorphism $X_1 \to X_2$. Choose $\chi \in C^N_0(X_1)$ and set*

$$Tu = \mathcal{F}^{-1}(\chi(\hat{u} \circ \Psi)).$$

Then T is bounded in $B_{\{c\}}$ with a norm which can be estimated in terms of the maximum of the derivatives of χ of order $\leq N$ and the derivatives of Ψ and Ψ^{-1} of order $\leq N+1$.

The above three lemmas are Theorems 14.1.4, 14.1.5, and 14.1.6 in [4].

For the spaces B_s, $s \geq 0$, we can choose $\{c\}_s = \{R_j^s\} = \{2^{(j-1)s}\}$ and $M_s = 2^s$. Then

$$2^{(j-2)s} = \frac{c_j}{M_s} \leq c_{j+1} = 2^{js} = M_s c_j.$$

In Lemma 3.1, we can choose $N = 2$ if $s = 1$. If $0 \leq s < 1$, we can choose $N = 1$. For the discussions in the rest of the paper, we always pick $N = 2$.

THEOREM 3.4. *Let $H(\xi) = H_1(\xi) + iH_2(\xi) \in C^3(\Omega)$, where Ω is an open set in $\mathbf{R}^n$; assume that* $\operatorname{Re}(\nabla H(\xi)) = \nabla H_1(\xi)$ *and* $\operatorname{Im}(\nabla H(\xi)) = \nabla H_2(\xi)$ *are linearly independent when $H(\xi) = 0$ in Ω, i.e., $|\nabla H(\xi)|_* \neq 0$ when $H(\xi) = 0$ in Ω. Then for fixed $\chi \in C_0^2(\Omega)$, there exists a constant C such that when $u \in B_1$ and $v \in B_0$,*

$$\left| \int \chi(\xi) H(\xi)^{-1} \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi \right| \leq C \|u\|_{B_1} \|v\|_{B_0} \tag{3.2}$$

and the constant C can be estimated in terms of the dimension n, the maximum of the derivatives of χ of order ≤ 2 and the derivatives of H of order ≤ 3 on $\operatorname{supp}(\chi)$, *the maximum of $|\nabla H(\xi)|_*^{-1}$ (see (1.8)) on* $\operatorname{supp}(\chi) \cap \bar{O}$ *(O is a neighborhood of $\{\xi \in \mathbf{R}^n : H(\xi) = 0\}$) and the maximum of $|H(\xi)|^{-1}$ on* $\operatorname{supp}(\chi) \backslash O$.

Proof. First, suppose that $\operatorname{supp} \chi$ is sufficiently small and $|\nabla H(\xi)|_* \neq 0$ on $\operatorname{supp} \chi$ such that for an open set $\Omega' \supset \operatorname{supp} \chi$ there is a C^3 diffeomorphism $\psi : \Omega'' \to \Omega'$, depending on H, $|\nabla H|_*$, and $|\nabla H|_*^{-1}$, with $H(\psi(\eta)) = \eta_1 + i\eta_2$. Then χ can be rewritten as $\chi = \chi_1 \chi_2$ for some functions $\chi_1, \chi_2 \in C_0^2(\Omega')$. Thus

$$\begin{aligned} &\int \chi_1(\xi) \chi_2(\xi) H(\xi)^{-1} \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi \\ &\quad = \int \frac{1}{\eta_1 + i\eta_2} \chi_1(\psi(\eta)) \hat{u}(\psi(\eta)) \chi_2(\psi(\eta)) \overline{\hat{v}(\psi(\eta))} |\det(\psi'(\eta))| d\eta. \end{aligned}$$

From Lemma 3.3 ($\psi \in C^3$ is used here to apply Lemma 3.3 since in our situation, $N = 2$), $\mathcal{F}^{-1}(\chi_1 \circ \psi \cdot \hat{u} \circ \psi) \in B_1$ and $\mathcal{F}^{-1}(\chi_2 \circ \psi \cdot \hat{v} \circ \psi) \in B_0$. Then Theorem 2.1 and Corollary 2.4 imply that

$$\mathcal{F}^{-1}\left(\frac{1}{\eta_1 + i\eta_2} \chi_1 \circ \psi \cdot \hat{u} \circ \psi \right) \in B_0^*.$$

Therefore,

$$\left| \int \chi_1(\xi) \chi_2(\xi) H(\xi)^{-1} \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi \right| \leq C \|u\|_{B_1} \|v\|_{B_0}.$$

In general, if $|\nabla H(\xi)|_* = 0$ at some points ξ in $\operatorname{supp}\chi$, write $\chi(\xi) = \chi_1(\xi) + \chi_2(\xi)$, where $|\nabla H(\xi)|_* \neq 0$ in $\operatorname{supp}\chi_1$ and $H(\xi) \neq 0$ in $\operatorname{supp}\chi_2$. Then it is clear that

$$\begin{aligned} \left| \int \chi_2(\xi) H(\xi)^{-1} \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi \right| &\leq C_1 \int |\hat{u}(\xi) \overline{\hat{v}(\xi)}| d\xi \\ &\leq C_1 \|u\|_{L_0^2} \|v\|_{L_0^2} \\ &\leq C_1 \|u\|_{B_1} \|v\|_{B_0}, \end{aligned}$$

with C_1 depending on $\max|\chi_2(\xi)|$ and the maximal lower bound of $|H(\xi)|$ on $\operatorname{supp}(\chi_2)$. On $\operatorname{supp}\chi_1$, using a partition of unity and the proof for small $\operatorname{supp}\chi$, we have

$$\left|\int \chi_1(\xi)H(\xi)^{-1}\hat{u}(\xi)\overline{\hat{v}(\xi)}d\xi\right| \leq C_2\|u\|_{B_1}\|v\|_{B_0},$$

where C_2 can be estimated in terms of the quantities stated in the theorem. Combining the estimates for χ_1 and χ_2 yields (3.2). □

Theorem 3.4 says that $Tu = \mathcal{F}^{-1}(\chi H^{-1}\hat{u})$ is bounded from B_1 to B_0^* and also from B_0 to B_1^*. Then by interpolation, we have the following result.

COROLLARY 3.5. *The operator $T_s u = \mathcal{F}^{-1}(\chi H^{-1}\hat{u})$, $u \in B_s$, is bounded from B_s to B_{1-s}^* for each $0 < s < 1$, that is,*

$$\|\mathcal{F}^{-1}(\chi H^{-1}\hat{u})\|_{B_{1-s}^*} \leq C_s\|u\|_{B_s}, \quad u \in B_s,$$

where c_s depends on n, s, and the quantities described in Theorem 3.4.

4. Estimate for complex simply characteristic polynomials. Let $P(\xi) = P_1(\xi) + iP_2(\xi)$ be a polynomial with constant complex coefficients, and assume that $|\nabla P(\xi)|_* \neq 0$ on $\{\xi \in \mathbf{R}^n : P(\xi) = 0\}$. Then for $\chi \in C_0^2(\mathbf{R}^n)$, we have shown in the previous section that

$$\|\mathcal{F}^{-1}(\chi P^{-1}\hat{f})\|_{B_{1-s}^*} \leq C_s\|f\|_{B_s}, \quad f \in B_s, \tag{4.1}$$

for $0 \leq s \leq 1$.

Now we want to impose a stronger condition on P to control the behavior of $|P(\xi)|$ at large ξ and thus allow an estimate like (4.1) without the cutoff function χ. A *simply characteristic* complex polynomial, defined in Definition 1.3, is exactly the kind of suitable condition that we want to impose, and it turns out to be an analogue of the one introduced by Agmon and Hörmander for real polynomials (Definition 1.1).

Now let us analyze condition (1.7). We may assume that there is no real vector $\eta \neq 0$ such that $P(\xi + t\eta) \equiv P(\xi)$ because if there is such a vector η, we can take it as a coordinate direction and obtain a polynomial in fewer variables. Our condition implies that $\tilde{P}_*(\xi) \to \infty$ as $\xi \to \infty$.

1. For each $\xi \in \mathbf{R}^n$, set $P_\xi(\eta) = P(\xi+\eta)/\tilde{P}_*(\xi)$. Then it is clear that the family $\{P_\xi(\eta) : \xi \in \mathbf{R}^n\}$ is uniformly bounded and equicontinuous on any compact set in $\mathbf{R}^n$. Thus by the Arzela–Ascoli theorem, the set M of all polynomials P_ξ and their limits form a compact set of polynomials. Straightforward calculations (the uniform boundedness of coefficients of all polynomials in M plays an essential role here; it allows us to interchange the limit operation with derivative operations and other operations involved in $|\nabla P_\xi(\eta)|_*$) give that

$$\partial_\eta^\alpha P_\xi(\eta) = \frac{P^{(\alpha)}(\xi+\eta)}{\tilde{P}_*(\xi)} \quad \text{for any multi-index } \alpha$$

and

$$|\nabla_\eta P_\xi(\eta)|_* = \frac{|\nabla P(\xi+\eta)|_*}{\tilde{P}_*(\xi)}.$$

If Q is any limit of P_ξ as $\xi \to \infty$, then

$$\partial_\eta^\alpha Q(\eta) = \lim_{\xi\to\infty} \frac{P^{(\alpha)}(\xi+\eta)}{\tilde{P}_*(\xi)}$$

and

$$|\nabla_\eta Q(\eta)|_* = \lim_{\xi\to\infty} \frac{|\nabla P(\xi+\eta)|_*}{\tilde{P}_*(\xi)}.$$

Therefore,

$$\tilde{Q}_*(\eta) = \lim_{\xi\to\infty} \frac{\sum_{|\alpha|\le m, |\alpha|\ne 1} |P^{(\alpha)}(\xi+\eta)| + |\nabla P(\xi+\eta)|_*}{\tilde{P}_*(\xi)},$$

and for any fixed $\theta \in \mathbf{R}^n$,

$$\begin{aligned} Q_\theta(\eta) &= \lim_{\xi\to\infty} \left[\left(\frac{P(\xi+\theta+\eta)}{\tilde{P}_*(\xi)} \right) \Big/ \left(\frac{\sum_{|\alpha|\le m, |\alpha|\ne 1} |P^{(\alpha)}(\xi+\theta)| + |\nabla P(\xi+\theta)|_*}{\tilde{P}_*(\xi)} \right) \right] \\ &= \lim_{\xi\to\infty} \frac{P(\xi+\theta+\eta)}{\sum_{|\alpha|\le m, |\alpha|\ne 1} |P^{(\alpha)}(\xi+\theta)| + |\nabla P(\xi+\theta)|_*} \\ &= \lim_{\xi+\theta\to\infty} P_{\xi+\theta}(\eta). \end{aligned} \tag{4.2}$$

This means that $Q_\theta(\eta)$ is also such a limit. In view of (1.7), we have

$$1 \le C(|Q(0)| + |\nabla Q(0)|_*) \tag{4.3}$$

if Q is any limit of P_ξ as $\xi \to \infty$. From (4.2), we also see that if $Q(\theta) = Q_1(\theta) + iQ_2(\theta) = 0$ at θ, then $|\nabla Q(\theta)|_* \ne 0$. This means that $\nabla Q_1(\theta)$ and $\nabla Q_2(\theta)$ are linearly independent on the zero set of Q.

2. Let z be any complex number in $\mathbf{C}$. Then condition (1.7) is valid for large $|\xi|$ if z_0 is replaced by z. Otherwise, we can find a limit Q with $|\nabla Q(0)|_* = 0$ and this contradicts (4.3). Thus the validity of (1.7) for large $|\xi|$ is independent of $z \in \mathbf{C}$ and means precisely that large real zeros ξ of $P(\xi) - z$ are *simple* in the sense that the real part $\nabla P_1(\xi)$ and the imaginary part $\nabla P_2(\xi)$ of $\nabla P(\xi)$ are linearly independent. For a *simply characteristic* complex polynomial, we conclude that (1.7) is uniformly valid when z belongs to a compact set that does not contain any *critical values* of P.

3. If $z \in \mathbf{C}$ is not a *critical value* of P, then when we set $P_\xi(\eta) = P(\xi+\eta) - z/\tilde{P}_*(\xi)$, condition (1.7) means that

$$P_\xi(0) = 0, \qquad |\nabla P_\xi(0)|_* \ge \frac{1}{C} \tag{4.4}$$

for $\xi \in \{\xi \in \mathbf{R}^n : P(\xi) - z = 0\}$. Combining (4.3) and (4.4), we see that the polynomials P_ξ and their limits as $\xi \to \infty$ either have a *simple* zero at $\eta = 0$ or do not equal 0 at $\eta = 0$. Furthermore, from (4.3), (4.4), and the uniform boundedness of the coefficients of all polynomials in M, we find that for some $r > 0$, independent of ξ, such that on $\Omega = \{|\eta| < r\}$, any Q in M either satisfies $|\nabla Q|_* \ne 0$ or else is uniformly bounded below.

Remark. If $P(\xi)$ is a hypoelliptic polynomial,

$$P^{(\alpha)}(\xi)/P(\xi) \to 0$$

when $\xi \to \infty$ in $\mathbf{R}^n$ for $\alpha \ne 0$. Thus hypoelliptic polynomials satisfy condition (1.7) for large ξ. Therefore if $z \in \mathbf{C}$ is not a *critical value* of P, condition (1.7) is satisfied for all $\xi \in \mathbf{R}^n$.

Theorem 4.1. *Assume that P is* simply characteristic *and let K be a compact subset of $\mathbf{C}$ containing no* critical values *of P. If $f \in B_1$, it follows that*

$$R(z)f = \mathcal{F}^{-1}((P(\cdot) - z)^{-1}\hat{f})$$

is in B_0^*, and we have the estimate

$$\|R(z)f\|_{B_0^*} \leq C \sup_{\xi \in \mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi)} \|f\|_{B_1} \tag{4.5}$$

for $z \in K$, where C depends only on the dimension n, the compact set K, and the constant c_P in condition (1.7).

Proof. It suffices to prove the theorem for some neighborhood $\tilde{K}$ of 0 in $\mathbf{C}$ when 0 is not a *critical value*. On account of the previous analysis and the proof of Theorem 3.4, we see that if $\chi \in C_0^\infty(\Omega)$, $\Omega = \{\eta \in \mathbf{R}^n : |\eta| < r\}$, there is a constant C (depending on the maximum of the derivatives of χ of order ≤ 2, a uniform bound of derivatives of all polynomials Q in M, a uniform bound of $|Q(\eta)|^{-1}$ on $\bar{\Omega}$ when $Q \neq 0$ in Ω, and a uniform bound of $|\nabla Q(\eta)|_*^{-1}$ on $\bar{\Omega}$ when Q has zeros in Ω) and a compact neighborhood K' of 0 in $\mathbf{C}$, both independent of ξ, such that

$$\left| \int |\chi(\eta)|^2 \left(\frac{1}{P_\xi - z/\tilde{P}_*(\xi)} \right) \hat{f}(\eta)\overline{\hat{g}(\eta)} d\eta \right| \leq C \|\mathcal{F}^{-1}(\chi \hat{f})\|_{B_1} \|\mathcal{F}^{-1}(\chi \hat{g})\|_{B_0}$$

if f, $g \in \mathcal{S}$, where $\mathcal{S}$ is the Schwartz space, and $z/\tilde{P}_*(\xi) \in K'$. Let $\tilde{K}$ be a neighborhood of 0 in $\mathbf{C}$ contained in $\tilde{P}_*(\xi)K'$ for all ξ. Making a translation of $\hat{f}$ and $\hat{g}$ and writing $\chi_\xi(\eta) = \chi(\eta - \xi)$, we have for any $z \in \tilde{K}$ that

$$\begin{aligned} &\left| \int \tilde{P}_*(\xi) \frac{1}{P(\eta) - z} \chi_\xi(\eta) \hat{f}(\eta) \overline{\chi_\xi(\eta)\hat{g}(\eta)} d\eta \right| \\ &\quad = \left| \int |\chi(\eta)|^2 \left(\frac{1}{P_\xi - z/\tilde{P}_*(\xi)} \right) \hat{f}(\eta + \xi) \overline{\hat{g}(\eta + \xi)} d\eta \right| \\ &\quad \leq C \|\mathcal{F}^{-1}(\chi \hat{f}(\xi + \cdot))\|_{B_1} \|\mathcal{F}^{-1}(\chi \hat{g}(\xi + \cdot))\|_{B_0} \\ &\quad \leq C \|\mathcal{F}^{-1}(\chi_\xi \hat{f})\|_{B_1} \|\mathcal{F}^{-1}(\chi_\xi \hat{g})\|_{B_0}. \end{aligned}$$

If we then write $\hat{f}_\xi = \chi_\xi \hat{f}$, $\hat{g}_\xi = \chi_\xi \hat{g}$, it follows that

$$\left| \int \frac{1}{P(\eta) - z} \chi_\xi(\eta) \hat{f}(\eta) \overline{\chi_\xi(\eta)\hat{g}(\eta)} d\eta \right| \leq C \sup_{\xi \in \mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi)} \|f_\xi\|_{B_1} \|g_\xi\|_{B_0}.$$

If we integrate with respect to ξ and use Lemma 4.2 below, we obtain

$$|(R(z)f, g)| \leq C \sup_{\xi \in \mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi)} \|f\|_{B_1} \|g\|_{B_0}. \tag{4.6}$$

From (4.3) and (4.4), we also see that a uniform bound of derivatives of all polynomials Q in M, a uniform bound of $|Q(\eta)|^{-1}$ on $\bar{\Omega}$ when $Q \neq 0$ in Ω, and a uniform bound of $|\nabla Q(\eta)|_*^{-1}$ on $\bar{\Omega}$ when Q has zeros in Ω can be estimated in terms of c_P in (1.7). Thus the constant C here depends only on the dimension n, the compact set K, and c_P in (1.7), and the proof is completed. □

LEMMA 4.2. *Let* $\chi \in C_0^\infty(\mathbf{R}^n)$ *and set* $\chi(D-\eta)u = \mathcal{F}^{-1}\chi(\cdot - \eta)\hat{u}$, $u \in \mathcal{S}'$. *Then we have*

$$\int \|\chi(D-\eta)u\|^2_{B_{\{c\}}} d\eta \le C_{M,\chi}\|u\|^2_{B_{\{c\}}}, \quad u \in B_{\{c\}},$$

for all $\{c\}$ *satisfying condition* (3.1).

This lemma is Theorem 14.1.7 in Hörmander's book [4].

From estimate (4.6), we see that $R(z)$ is bounded from B_0 to B_1^* as well. Then by interpolation, we obtain that $R(z)$ is also bounded from B_s to B_{1-s}^* for $0 < s < 1$. This completes the proof of Theorem 1.2.

Remark.

1. If $\{P(\xi,\zeta)\}$ is a family of polynomials of $\xi \in \mathbf{R}^n$ depending on a parameter ζ in a subset $M_\zeta \subset \mathbf{C}^n$ and condition (1.7) is valid for all $\zeta \in M_\zeta$, with the constant $c_P(\zeta)$ depending on ζ, then

$$|(R(z,\zeta)f,g)| \le C(s,c_P(\zeta)) \sup_{\xi\in\mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi,\zeta)}\|f\|_{B_s}\|g\|_{B_{1-s}} \tag{4.7}$$

for all $f \in B_s$ and $g \in B_{1-s}$ $(0 \le s \le 1)$ with $C(s,c_P(\zeta))$ dependent on ζ.

2. If c_P is independent of $\zeta \in M_\zeta$, then

$$|(R(z,\zeta)f,g)| \le C(s,c_P) \sup_{\xi\in\mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi,\zeta)}\|f\|_{B_s}\|g\|_{B_{1-s}} \tag{4.8}$$

for all $f \in B_s$ and $g \in B_{1-s}$ $(0 \le s \le 1)$ with $C(s,c_P)$ independent of ζ.

These two remarks will be useful in constructing exponentially growing solutions of a partial differential equation $(P(D+\zeta)-\lambda)u = f$ to study inverse problems. Some examples can be found in [8].

5. The uniqueness of solutions and the behavior at infinity. In this section, we use the estimate in Theorem 1.2 to discuss the uniqueness and the asymptotic behavior in an average sense of solutions of the partial differential equation

$$P(D)u = f \in B_s, \quad 0 \le s \le 1. \tag{5.1}$$

We assume that the symbol $P(\xi) = P_1(\xi) + iP_2(\xi)$ of $P(D)$ is *simply characteristic* and 0 is not a *critical value*. Then from Theorem 1.2, we have

$$R(0)f = \mathcal{F}^{-1}((P(\cdot))^{-1}\hat{f}) \in B_{1-s}^*$$

with the bound

$$\|R(0)f\|_{B_{1-s}^*} \le C(s,c_P) \sup_{\xi\in\mathbf{R}^n} \frac{1}{\tilde{P}_*(\xi)}\|f\|_{B_s}$$

for $0 \le s \le 1$.

THEOREM 5.1. *Let* $P(\xi) = P_1(\xi) + iP_2(\xi)$ *be a* simply characteristic *complex polynomial and let* 0 *not be a* critical value *of* P. *Suppose that the zero set* M *of* $P(\xi)$ *is a* C^1 *submanifold of codimension* 2. *Then the solution to* (5.1) *in* B_{1-s}^* *is unique if* $0 < s \le 1$ *and* $u = \mathcal{F}^{-1}(P(\cdot)^{-1}\hat{f})$. *Moreover, if* $f \in B_s$ *and* $0 < s < 1$, *then*

$$\lim_{R\to\infty} \sup \frac{1}{R^{2(1-s)}} \int_{|x|<R} |u(x)|^2 dx = 0.$$

In particular, if $f \in B_1$, *then*

$$\lim_{R\to\infty} \sup \frac{1}{R^{2(1-s)}} \int_{|x|<R} |u(x)|^2 dx = 0$$

for all $0 < s < 1$.

Proof. To prove uniqueness, we need the following lemma, proved in [2].

LEMMA 5.2. *Let* $u \in \mathcal{S}' \cap L^2_{\text{loc}}$ *and assume that*

$$\lim_{R\to\infty} \sup \frac{1}{R^k} \int_{|x|<R} |u(x)|^2 dx < \infty.$$

If the restriction of the Fourier transform $\hat{u}$ *to an open subset* Ω *of* $\mathbf{R}^n$ *is supported by a* C^1 *submanifold* M *of codimension* k, *then it is an* L^2 *density* $\hat{u}_0 ds$ *on* M *and*

$$\int_M |\hat{u}_0|^2 ds \leq c \lim_{R\to\infty} \sup \frac{1}{R^k} \int_{|x|<R} |u(x)|^2,$$

where c *depends only on the dimension* n.

Suppose that we have u_1 and u_2 in B^*_{1-s} satisfying equation (5.1). Set $u = u_1 - u_2$. Then

$$P(D)u = 0.$$

This implies that $\hat{u}$ is supported on $M = \{\xi \in \mathbf{R}^n : P(\xi) = 0\}$, which is a C^1 submanifold of codimension 2. Since $u \in B^*_{1-s}$, $(0 < s \leq 1)$,

$$\begin{aligned}
&\lim_{R\to\infty} \sup \frac{1}{R^2} \int_{|x|<R} |u(x)|^2 dx \\
&\quad = \lim_{R\to\infty} \sup \frac{1}{R^{2s}} \frac{1}{R^{2(1-s)}} \int_{|x|<R} |u(x)|^2 dx \\
&\quad = \lim_{R\to\infty} \frac{1}{R^{2s}} \lim_{R\to\infty} \sup \frac{1}{R^{2(1-s)}} \int_{|x|<R} |u(x)|^2 dx \\
&\quad = 0.
\end{aligned}$$

The last equality follows from

$$\sup_{R\geq 1} \frac{1}{R^{2(1-s)}} \int_{|x|<R} |u(x)|^2 dx \leq \frac{2^{2(1-s)}}{(1-2^{-2(1-s)})} \|u\|^2_{B^*_{1-s}} < \infty.$$

By Lemma 5.2,

$$\int_M |\hat{u}|^2 ds \leq c \lim_{R\to\infty} \sup \frac{1}{R^2} \int_{|x|<R} |u(x)|^2 dx = 0.$$

This proves uniqueness.

To prove that the limit is 0, we note that if $0 < s < 1$ and $f \in B_s$, $R(0)f \in L^2_{-1+s}$ (see the proof of Corollary 2.5), that is,

$$\int_{\mathbf{R}^n} (1+|x|^2)^{-1+s} |\mathcal{F}^{-1}(P(\cdot)^{-1}\hat{f})(x)|^2 dx < \infty,$$

then

$$\lim_{j\to\infty}\sup\int_{x\in\Omega_j}(1+|x|^2)^{-1+s}|\mathcal{F}^{-1}(P(\cdot)^{-1}\hat{f})(x)|^2dx=0.$$

This implies that

$$\lim_{j\to\infty}\sup\frac{1}{R_j^{2(1-s)}}\int_{x\in\Omega_j}|\mathcal{F}^{-1}(P(\cdot)^{-1}\hat{f})(x)|^2dx=0.$$

For any $\epsilon>0$, there is an integer $N>0$ such that when $m>N$,

$$\frac{1}{R_m^{2(1-s)}}\int_{\Omega_m}|u(x)|^2dx<\epsilon,$$

where $u(x)=\mathcal{F}^{-1}(P(\cdot)^{-1}\hat{f})(x)$. Then

$$\begin{aligned}\frac{1}{R_m^{2(1-s)}}\int_{|x|<R_m}|u(x)|^2dx &= \frac{1}{R_m^{2(1-s)}}\sum_{j=1}^{m}\int_{\Omega_j}|u(x)|^2dx\\ &=\sum_{j=1}^{m}\frac{1}{(2^{2(1-s)})^{m-1}}\int_{\Omega_j}|u(x)|^2dx\\ &=\sum_{j=1}^{m}\frac{1}{(2^{2(1-s)})^{m-j}}\frac{1}{(2^{2(1-s)})^{j-1}}\int_{\Omega_j}|u(x)|^2dx\\ &=\sum_{j=1}^{N}\frac{1}{(2^{2(1-s)})^{m-j}}\frac{1}{R_j^{2(1-s)}}\int_{\Omega_j}|u(x)|^2dx\\ &\quad+\sum_{j=N+1}^{m}\frac{1}{(2^{2(1-s)})^{m-j}}\frac{1}{R_j^{2(1-s)}}\int_{\Omega_j}|u(x)|^2dx\\ &\le\|u\|^2_{B^*_{1-s}}\sum_{j=1}^{N}\frac{1}{(2^{2(1-s)})^{m-j}}+\epsilon\sum_{j=N+1}^{m}\frac{1}{(2^{2(1-s)})^{m-j}}\\ &\le\|u\|^2_{B^*_{1-s}}\frac{1}{(2^{2(1-s)})^{m-N}}\frac{1}{1-2^{-2(1-s)}}+\epsilon\frac{1}{1-2^{-2(1-s)}}\\ &\le\frac{\epsilon}{1-2^{-2(1-s)}}(1+\|u\|^2_{B^*_{1-s}})\end{aligned}$$

for $m>N$ large enough. Therefore,

$$\begin{aligned}&\lim_{R\to\infty}\sup\frac{1}{R^{2(1-s)}}\int_{|x|<R}|u(x)|^2dx\\ &\quad=\lim_{R\to\infty}\sup_{R'\ge R}\frac{1}{(R')^{2(1-s)}}\int_{|x|<R'}|u(x)|^2dx=0. \qquad\square\end{aligned}$$

Acknowledgments. This paper is part of my Ph.D. dissertation at the University of Rochester. I am greatly indebted to my Ph.D. advisor, Professor Adrian Nachman, for his guidance and many helpful ideas. It is also a pleasure to thank Professors Richard Lavine and Allan Greenleaf for their useful suggestions. I would also like to thank the referees for spending time reading my paper carefully and giving some useful suggestions.

REFERENCES

[1] S. Agmon, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 151–218.

[2] S. Agmon and L. Hörmander, *Asymptotic properties of solutions of differential equations with simple characteristics*, J. Anal. Math., 30 (1976), pp. 1–37.

[3] R. Beals and R. R. Coifman, *Multidimensional Inverse Scattering and Nonlinear PDE*, Proc. Sympos. Pure Math. 43, AMS, Providence, RI, 1985, pp. 45–70.

[4] L. Hörmander, *The Analysis of Linear Partial Differential Operators* II, Springer-Verlag, Berlin, 1983.

[5] V. Isakov, *Completeness of products of solutions and some inverse problems for PDE*, J. Differential Equations, 92 (1991), pp. 305–317.

[6] M. Ikehata, *A special Green function for the biharmonic operator and its application to an inverse boundary value problem*, J. Comput. Math. Appl., 22 (1991), pp. 53–66.

[7] R. Lavine and A. Nachman, *Multidimensional Inverse Problems for Singular Potentials*, in preparation.

[8] C. Liu, *Sharp estimates for solutions of P.D.E. and uniqueness for a general class of inverse problems*, Ph.D. thesis, University of Rochester, Rochester, NY, 1994.

[9] A. Nachman, *Reconstructions from boundary measurements*, Ann. Math., 128 (1988), pp. 531–576.

[10] A. Nachman, *Inverse scattering at fixed energy*, in Proc. 10th International Congress on Mathematical Physics, K. Schmüdgen, ed., Springer-Verlag, Berlin, 1992, pp. 434–441.

[11] A. I. Nachman and M. J. Ablowitz, *A multidimensional inverse scattering method*, Stud. Appl. Math., 71 (1984), pp. 243–250.

[12] R. Novikov and G. Henkin, *$\bar{\partial}$-equation in the multidimensional inverse scattering problems*, Uspekhi Mat. Nauk., 42 (1987), pp. 93–152.

[13] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 944–970, July 1997

MULTIFRACTAL FORMALISM FOR FUNCTIONS PART I: RESULTS VALID FOR ALL FUNCTIONS*

S. JAFFARD†

Abstract. The multifractal formalism for functions relates some functional norms of a signal to its "Hölder spectrum" (which is the dimension of the set of points where the signal has a given Hölder regularity). This formalism was initially introduced by Frisch and Parisi in order to numerically determine the spectrum of fully turbulent fluids; it was later extended by Arneodo, Bacry, and Muzy using wavelet techniques and has since been used by many physicists. Until now, it has only been supported by heuristic arguments and verified for a few specific examples. Our purpose is to investigate the mathematical validity of these formulas; in particular, we obtain the following results:

- The multifractal formalism yields for any function an upper bound of its spectrum.
- We introduce a "case study," the *self-similar functions*; we prove that these functions have a concave spectrum (increasing and then decreasing) and that the different formulas allow us to determine either the whole increasing part of their spectrum or a part of it.
- One of these methods (the wavelet-maxima method) yields the complete spectrum of the self-similar functions.

We also discuss the implications of these results for fully developed turbulence.

Key words. multifractal formalism, self-similarity, wavelet transform

AMS subject classifications. Primary, 26A15; Secondary, 76F99

PII. S0036141095282991

1. Introduction and statement of results.

One-dimensional multifractal measures have been the object of many investigations by mathematicians and theoretical physicists (see, for instance, [5], [7], [12], [23], and the references therein). Basically, such measures have very different "scalings" from point to point, i.e., for such a measure μ, if I is an interval, the quantity $\mu(I)$ scales like $|I|^\alpha$, where the exponent α differs very much following the position of the center of the interval I. Such measures are important because they are natural measures carried by some strange attractors and thus appear in the modeling of many natural phenomena (diffusion-limited aggregates, invariant measures of dynamical systems, voltage drop across a random transistor network, etc.; see [2] and the references therein).

It may happen that the natural, fractal-like object that one wants to understand is not a set or a measure but a function. The study of multifractal functions has proved important in several domains of physics. Examples include plots of random walks, interfaces developing in reaction-limited growth processes, and turbulent velocity signals at inertial range (see [3]). The relevant mathematical tool studied in this context is the *Hölder spectrum*, also frequently called *spectrum of singularities*; this function associates with each positive α the Hausdorff dimension of the set where F is approximately Hölder of order α (in a sense to be made precise). The most important example where one would like to determine the spectrum of singularities of a function is the velocity of fully developed turbulence. The reason is that turbulent flows are not spatially homogeneous: the irregularity of the velocity seems to differ widely from point to point. This phenomenon, called "intermittency," suggests that the determination of the Hölder spectrum of the velocity of the fluid might be

[*] Received by the editors March 13, 1995; accepted for publication (in revised form) April 5, 1996. http://www.siam.org/journals/sima/28-4/28299.html

[†] CMLA, École Normale Supérieure Cachan, 61 Avenue du Président Wilson, 94235 Cachan cedex, France and Département de Mathématiques, Faculté des Sciences et Technologie, Université Paris XII, 61 Avenue du Gal. de Gaulle, 94010 Créteil cedex, France (jaffard@cmla.ens-cachan.fr).

a nontrivial function and thus would yield important information on the nature of turbulence.

The first problem in this ambitious program is the numerical determination of the spectrum. Obviously, it is almost impossible to deduce it from the mathematical definition since it involves the successive determination of several intricate limits. The only method is to find some "reasonable" assumptions under which the spectrum could be derived using only "averaged quantities" (which should be numerically stable) extracted from the signal. Such formulas for the spectrum can be guessed heuristically using similarities with statistical physics. Frisch and Parisi [14] proposed, in one dimension, a formula using the L^p modulus of continuity of the velocity along one axis. Arneodo, Bacry, and Muzy (in [2], [3], and [26]) proposed, also in one dimension, other formulas based on the wavelet transform of the signal, and they proved their formulas' validity when the function considered is the indefinite integral of a multinomial measure or a C^∞ perturbation of such a measure. The origin of this method may be traced to the seminal work of Mandelbrot [23], and it has been used a great deal by physicists (see for instance [4], [12], [24], and the references therein), so the scope of its mathematical validity has become an important issue.

Our purpose in this paper is twofold:

- In Part I, we give some general results concerning the multifractal formalism. We show that for any function, it yields an upper bound of its Hölder spectrum, but we also show via some explicit counterexamples that, in general, it does not yield the exact spectrum.

- In Part II, we introduce and study a model case, "self-similar functions," and prove that the multifractal formalism holds for these functions. Examples of such functions include the indefinite integrals of self-similar measures, but they also include widely oscillating, several-dimensional functions—two requirements which are obviously needed, for instance, in any realistic model of turbulence.

Before describing the multifractal formalism, we need to recall some definitions and notation concerning the Hölder regularity of functions.

Suppose that α is a positive real number; a function $F : \mathbb{R}^m \to \mathbb{R}$ is $C^\alpha(x_0)$ if there exists a polynomial P of degree less than α such that

$$|F(x) - P(x - x_0)| \leq C|x - x_0|^\alpha \tag{1.1}$$

and F belongs to $\Gamma^\alpha(x_0)$ if (see [18])

$$\begin{cases} \forall\beta > \alpha, & F \notin C^\beta(x_0), \\ \forall\beta < \alpha, & F \in C^\beta(x_0). \end{cases}$$

A function F is C^α (or $C^\alpha(\mathbb{R}^m)$) if (1.1) holds for any x in $\mathbb{R}^m$, the constant C being uniform. (Using *this* definition C^1 means Lipschitz.) We also need the two following definitions which assert (in two slightly different ways) that the singularity of F at x_0 can be measured on a "large" set near x_0. We denote by mes$A$ the Lebesgue measure of a set A.

Definition 1.1. *Let $\alpha > -m$; a point x_0 is a strong α-singularity of F if there* *exist* $C, C' > 0$ *such that* $\forall P$ *polynomial of degree at most* α, $\forall j, \exists A_j, B_j$,

$$\text{(1.2)} \quad \begin{cases} \text{mes} A_j \geq C2^{-mj}, \qquad \text{mes} B_j \geq C2^{-mj}, \\ \forall x \in A_j \cup B_j, \quad |x - x_0| \leq 2^{-j}, \\ \forall x \in A_j, \forall y \in B_j, \quad (F(x) - P(x - x_0)) - (F(y) - P(y - x_0)) \geq C'2^{-\alpha j}. \end{cases}$$

Note that if $\alpha < 1$, the last condition reduces to $F(x) - F(y) \geq C'2^{-\alpha j}$. The *wavelet transform* of a function F is defined as follows:

$$C(a, b) = \frac{1}{a^m} \int F(t) \psi\left(\frac{t - b}{a}\right) dt,$$

where ψ is a radial function with moments of order less than K vanishing and with derivatives of order less than K having fast decay (with a K "large enough" depending on the properties of F that we want to analyze).

Definition 1.2. *A point* x_0 *is a wavelet* α*-singularity of* F *if there exist wavelet coefficients* $C(a_n, b_n)$ *in a cone pointing towards* x_0 *(i.e.,* $\mid b_n - x_0 \mid \leq Ca_n$*) such that* $a_n \to 0$, $a_n/a_{n+1} \leq C$, *and*

$$\text{(1.3)} \qquad |C(a_n, b_n)| \geq Ca_n^\alpha.$$

We will prove in section 2 that the two previous definitions are related and that if F is $C^\alpha(x_0)$ and x_0 is a wavelet α-singularity of F, then x_0 is a strong α-singularity of F.

We can now define the object of our study.

Definition 1.3. *The Hölder spectrum of a function* F *is the function* $d(\alpha)$ *defined for each* $\alpha \geq 0$ *as follows:*

$d(\alpha)$ *is the Hausdorff dimension of the set of points* x_0 *where* F *belongs to* $\Gamma^\alpha(x_0)$.

Remark. We will sometimes also call the function $D(\alpha)$, which is the packing dimension of the strong α-singularities, the *packing dimension spectrum.*

The two definitions of dimension that we use will be recalled when needed. Note that $d(\alpha)$ and $D(\alpha)$ are defined point by point. We will consider mainly $d(\alpha)$ except in section 4 of Part I and section 6 of Part II.

We are now in a position to describe the methods used by Frisch and Parisi on one side and Arneodo, Bacry, and Muzy on the other in order to determine the spectrum of singularities of functions.

- The *structure function* method first requires the computation of

$$S_q(l) = \int_{\mathbb{R}^m} |F(x + l) - F(x)|^q dx.$$

Assuming that the order of magnitude of $S_q(l)$ is $|l|^{\zeta(q)}$ when $l \to 0$, the Hölder spectrum is computed using the formula

$$\text{(1.4)} \qquad d(\alpha) = \inf_q (q\alpha - \zeta(q) + m).$$

(We will define $\zeta(q)$ precisely below.)

- In the *wavelet-transform integral* method, one computes

$$\tilde{Z}(a, q) = \int_{\mathbb{R}^m} |C(a, b)|^q db,$$

and then if the order of magnitude of $\tilde{Z}(a,q)$ is $a^{\eta(q)}$,

$$d(\alpha) = \inf_q (q\alpha - \eta(q) + m). \tag{1.5}$$

• In order to describe the *wavelet-transform maxima* method, we first have to introduce the notion of a *line of maxima*; consider for a given $a' > 0$ the local maxima of the function $b \to C(a', b)$; generically, they belong to a line of maxima $b = l(a)$ defined in a small left-neighborhood $[a'', a']$ of a' by the condition that $b \to C(a,b)$ has a local maximum for $b = l(a)$. Usually, one cannot choose $a'' = 0$ because the lines of maxima have ramifications called "fingerprints." The wavelet-transform maxima method first requires the computation of

$$Z(a,q) = \sum_l \sup_{(b=l(a))} |C(a,b)|^q, \tag{1.6}$$

where l is a line of maxima of the wavelet transform defined on $[a'', a']$ and where the sum is taken on all lines of local maxima defined in left-neighborhoods $[a'', a']$ of a'. If the order of magnitude of $Z(a,q)$ is $a^{\theta(q)}$, then

$$d(\alpha) = \inf_q (qh - \theta(q)). \tag{1.7}$$

Numerically, according to [3], the most reliable method seems to be the last one, probably because the restriction of the computation to the maxima insures that small errors are less likely to be taken into account since at the maxima, they are relatively less important. More generally, methods that involve the wavelet transform are numerically more stable, probably because they involve only averaged quantities and not the direct values of the function. The use of such quantities has been advocated in [15]. The structure function method involves only order-one differences so that it is clearly unfit for computing the spectrum $d(\alpha)$ when α is larger than 1.

Since the scalings assumed above do not necessarily hold, we use the following definitions. Let

$$\zeta(q) = \liminf_{l \to 0} \frac{\log S_q(l)}{\log |l|}; \tag{1.8}$$

$$\eta(q) = \liminf_{a \to 0} \frac{\log \tilde{Z}(a,q)}{\log a}; \tag{1.9}$$

$$\theta(q) = \liminf_{a \to 0} \frac{\log Z(a,q)}{\log a}, \tag{1.10}$$

The multifractal formalism may seem surprising at first glance because it relates pointwise behaviors to global estimates. Before giving some mathematical explanations for it, it may be enlightening to give the heuristic classical argument from which it is derived. Although this argument cannot be transformed into a correct mathematical proof, it at least shows why these formulas can be expected to hold, and a careful study of its defects shows under which type of additional conditions it should be mathematically correct.

We calculate the contribution of singularities of order α to the integral

$$\int_{\mathbb{R}^m} |F(x+l) - F(x)|^q dx.$$

Near a singularity of order α, we have, in a small box of size $|l|$,

$$|F(x+l)-F(x)|^q \sim |l|^{\alpha q}.$$

If the dimension of these singularities is $d(\alpha)$, it means that there are about $|l|^{-d(\alpha)}$ such boxes, each of volume $|l|^m$, so that the total contribution to the integral is $|l|^{\alpha q+m-d(\alpha)}$. The real order of magnitude of the integral is given by the largest contribution, which, since $l \to 0$ is given by the smallest exponent, is such that

$$\zeta(q) = \inf_{\alpha}(\alpha q + m - d(\alpha)). \tag{1.11}$$

This formula is not the one that we are looking for since we know $\zeta(q)$ and are looking for $d(\alpha)$, but if it holds and if d is concave (we will see that in general this assumption need not be verified; however in many cases it is), $d(\alpha)$ is recovered by an inverse Legendre transform formula which yields (1.4). Of course, if $d(\alpha)$ is not concave, one expects the right-hand side of (1.4) to yield only the convex hull of the spectrum.

In all cases, (1.11) is more likely to hold because the concavity problem does not appear there. (A straightforward application of Young's formula shows that $\zeta(q)$ is always concave.)

In the first part of this paper, the following results will be proved.

THEOREM 1.4. *If $q > 1$ and $\zeta(q) < q$, then $\zeta(q) = \eta(q)$ for any function F. In general, these functions need not be related to $\theta(q)$.*

If F is a function of one real variable, and $0 < \eta(1) < 1$, the box dimension of the graph of F is $2 - \eta(1)$.

The following upper bound holds for any function F such that $\eta(p) > m$ $\forall p$:

$$d(\alpha) \le \inf_{p}(m - \eta(p) + \alpha p). \tag{1.12}$$

Also, without any assumption on η,

$$D(\alpha) \le \inf_{p}(m - \eta(p) + \alpha p).$$

In general, (1.12) *cannot be an equality*; *more precisely, let $d(\alpha)$ be a Riemann-integrable positive function on $\mathbb{R}^+$. There exists F_1 and F_2 which share the same function η, but the spectrum of F_1 is $d(\alpha)$ and F_2 is C^∞ except at the origin* (*so that its spectrum is equal to $-\infty$ everywhere except at one point*).

Some counterexamples will show that a smooth function (with a large $\eta(p)$) may nonetheless be such that $\theta(p)$ can be arbitrarily small. (The case $\theta(p) = -\infty$ $\forall p > 0$ can even happen.) The wavelet-transform maxima method need not be correct, even in the more precise framework of self-similar functions, where the other methods will work. However, after a slight modification, it yields the correct spectrum for self-similar functions. The mathematical problem with using (1.6) is that the lines of maxima can be too close to each other. In that case, we should instead keep for each interval of width a only one line passing through this interval that yields the largest contribution. However, the reader will see that the mathematical counterexamples where $\eta(q) \ne \theta(q)$ are very contrived, and the author's belief is that for practical applications, (1.5) and (1.7) have the same range of validity.

The last assertion in the theorem is stronger than the mere failure of the Legendre transform formulas. It asserts that there is not enough information in the function η to determine the spectrum. In particular, contrary to a common belief, the fact that

η is not linear does not imply that the signal has a multifractal structure. It also shows that, mathematically, "any" function $d(\alpha)$ can be a spectrum. It is surprising to notice that in several different fields of application, this does not seem to be the case. The spectra computed numerically have always the same shape—roughly speaking, the upper part of an ellipse. This is actually the shape we will find for self-similar functions. There can be several explanations to this analogy. Either (a) these physical signals satisfy some "scaling-invariance" properties which makes them fit in the framework (perhaps generalized in some ways) of self-similar functions or (b) a pessimistic explanation could be that, these spectra being (perhaps wrongly) calculated using a Legendre transform, the convex hull of the true spectrum is actually calculated and not the spectrum itself—hence this "generic" concave shape. We will also see that these examples answer the following problem raised by Daubechies and Lagarias in [9], which is somehow converse to the multifractal formalism: Is η the Legendre transform of $m-d(\alpha)$? Positive answers to this problem find fewer applications than the multifractal formalism since in practice one wants to obtain $d(\alpha)$ knowing $\eta(p)$ or $\zeta(p)$ and not the converse; nonetheless, it might hold more generally (see [9]). The problem raised by Daubechies and Lagarias is to find explicit counterexamples. We will see that in most cases, F_1 and F_2 are such counterexamples.

One of the referees of this paper raised the problem of a relationship between $\theta(q)$ and $\eta(q)$ such as

$$\theta(q) \leq \eta(q) - m.$$

This is true for self-similar functions satisfying the closed-set condition because then the regions where the wavelet transform is large (and these are the regions taken into account to estimate $\eta(q)$) are isolated so that there must exist a local maximum of the wavelet transform in the neighborhood. In general, however, we have no answer to this problem.

We now define self-similar functions by analogy with self-similar sets.

Recall that a set K is *strictly self-similar* if it is a finite union of disjoint subsets $K_1, \dots, K_d$ which can be deduced from K by similitudes. For instance, the triadic Cantor set and the Van Koch curve are self-similar. These sets have been widely studied, as have the measures supported on them. They play an important role in the modeling of several physical phenomena (see, for instance, [5], [16], [12], and [3]).

Suppose that F is continuous and compactly supported and let Ω be the bounded open subset of $\mathbb{R}^n$ such that $\bar{\Omega} = \operatorname{supp}(F)$. The intuitive idea of a self-similar function is that there should exist disjoint subsets $\Omega_1, \dots, \Omega_d$ of Ω such that the graph of F restricted to each Ω_i is a "contraction" of the graph of F, modulo a certain error, which is supposed to be smooth. First, suppose that "smooth" means Lipschitz, and let us formalize this definition.

There should exist similitudes $(S_i)_{i=1,\dots,d}$ such that if $S_i(\Omega) = \Omega_i$,

$$\begin{aligned} &\forall i, \quad \Omega_i \subset \Omega, \\ &\Omega_i \cap \Omega_j = \phi \text{ if } i \neq j, \\ \forall x \in \Omega_j, \quad &F(x) = \lambda_j F(S_j^{-1}(x)) + g_j(x) \quad \text{with } g_j \text{ Lipschitz on } \bar{\Omega}_j. \end{aligned} \tag{1.13}$$

We suppose that S_i are *contractions*, i.e., the product of an isometry by the mapping $x \to \mu_i x$, where $\mu_i < 1$.

Equation (1.13) does not tell how F behaves outside Ω_i. We make the assumption that it is smooth, i.e., Lipschitz, outside $\bigcup \Omega_i$.

Since $F(S_j^{-1}(x)) = 0$ if $x \notin \Omega_j$,

$$F(x) = \sum_{i=1}^{d} \lambda_j F(S_j^{-1}(x)) + g(x),$$

where $g = g_j$ on Ω_j, $g = F$ outside $\bigcup \Omega_j$, and g is obviously continuous since F is continuous; since it is Lipschitz on $\bigcup \bar{\Omega}_j$ and outside $\bigcup \Omega_j$, g is uniformly Lipschitz.

This equation holds for any Lipschitz function F (use it as a definition for g when all $\lambda_j = 0$) so that it is interesting only if F is not uniformly Lipschitz, and in that case, we will be interested in determining the points where F is C^α for $\alpha < 1$.

We will generalize this model slightly by assuming that g is $C^k(\mathbb{R}^m)$ and not necessarily compactly supported but also that the derivatives of g of order less than k have fast decay. The same remark shows that in this case, we should suppose that F is not $C^k(\mathbb{R}^m)$, and we will be interested in determining where F is C^α for $\alpha < k$. We will thus use the following definition.

Definition 1.5. *A function $F : \mathbb{R}^m \to \mathbb{R}$ is self-similar (of order $k \in \mathbb{R}^+$) if the three following conditions hold:*

- *There exists a bounded open set Ω and $S_1, \dots, S_d$ contractive similitudes such that*

$$S_i(\Omega) \subset \Omega, \tag{1.14}$$

$$S_i(\Omega) \cap S_j(\Omega) = \emptyset \quad \text{if } i \neq j. \tag{1.15}$$

(The S_i's are the product of an isometry with the mapping $x \to \mu_i x$, where $\mu_i < 1$.)

- *There exists a C^k function g such that g and its derivatives of order less than k have fast decay and F satisfies*

$$F(x) = \sum_{i=1}^{d} \lambda_i F(S_i^{-1}(x)) + g(x). \tag{1.16}$$

- *The function F is not uniformly C^k in a certain closed subset of Ω.*

Recall that g has fast decay if

$$\forall n \in \mathbb{N}, \quad |g(x)| \leq \frac{C_n}{(1+|x|)^n}.$$

Let

$$\alpha_{\min} = \inf_{i=1,\dots,d} \left(\frac{\log \lambda_i}{\log \mu_i} \right), \qquad \alpha_{\max} = \sup_{i=1,\dots,d} \left(\frac{\log \lambda_i}{\log \mu_i} \right).$$

We use this notation because $\alpha_{\min}$ will turn out to be the smallest pointwise Hölder regularity exponent of F and $\alpha_{\max}$ the largest (lower than k). Let τ be the function defined by

$$\sum_{i=1}^{d} \lambda_i^a \mu_i^{-\tau(a)} = 1.$$

Some results concerning the multifractal formalism for self-similar functions are summed up in the following theorem and will be proved in the second part of the paper.

Theorem 1.6. *Suppose that F is self-similar. If $\alpha_{\min} > 0$, the function $d(\alpha)$ vanishes outside $[\alpha_{\min}, \alpha_{\max}] \cup [k, +\infty)$ and is analytic and concave on $[\alpha_{\min}, \alpha_{\max}]$. Its maximal value $d_{\max}$ on this interval satisfies*

$$\sum \mu_i^{d_{\max}} = 1.$$

Let α_0 be the value for which this maximum is attained. First, suppose that g is C^∞. If $\alpha \leq \alpha_0$, $d(\alpha)$ can be obtained by computing the Legendre transform of either $\eta(q) - m$ or $\zeta(q) - m$.

If g is only C^k, let p_0 be defined by $\eta(p_0) = kp_0$ and let $\alpha_1 < \alpha_0$ be the value of the inverse Legendre transform of $\eta(q) - m$ at p_0; if $\alpha \leq \alpha_1$, $d(\alpha)$ can be obtained by computing the Legendre transform of either $\eta(q) - m$ or $\zeta(q) - m$.

Without any assumption on $\alpha_{\min}$, if $\sum \mid \lambda_j \mid \mu_j^m < 1$, the same results hold if we replace $d(\alpha)$ by $D'(\alpha)$, the packing dimension of the wavelet α-singularities (or by $D(\alpha)$ if g and λ_i are positive and if furthermore the separated open-set condition holds).

We will also prove that in some cases, the wavelet-maxima method can be modified so that it yields the whole spectrum of self-similar functions (see Theorem 2.2 in Part II).

Corollary 8.5 in Part II of this paper will extend this result to a larger class of functions than self-similar functions.

Before we begin to study the multifractal formalism for functions, we show its relationship to the multifractal formalism for measures. We recall that if μ is a probability measure on $[0, 1]$, one defines

$$\tau(q) = \lim_{j \to +\infty} \frac{1}{j \log 2} \log \sum \left(\mu\left(\left[\frac{k}{2^j}, \frac{k+1}{2^j} \right] \right) \right)^q$$

and

$$E_\alpha = \left\{ t : \frac{\log \mu(I_n(t))}{\log |I_n(t)|} \to \alpha \right\},$$

where $I_n(t)$ is the interval $[k/2^j, (k+1)/2^j]$ which contains t. The multifractal formalism for measures asserts that the dimension of E_α is the Legendre transform of τ (see, for instance, [5] and [12] for mathematical results concerning this assertion). Let F be an indefinite integral of μ ($F(x) = \mu([0, x])$). Clearly,

$$t \in E_\alpha \Leftrightarrow |F(x+h) - F(x)| \sim h^\alpha$$

and

$$\sum \left(\mu\left(\left[\frac{k}{2^j}, \frac{k+1}{2^j} \right] \right) \right)^q \sum \left| F\left(\frac{k}{2^j} \right) - F\left(\frac{k+1}{2^j} \right) \right|^q \sim 2^j \int |F(x+2^{-j}) - F(x)|^q dx.$$

Thus if F is the indefinite integral of a probability measure supported on $[0, 1]$, the two multifractal formalisms are identical. However, in dimensions larger than one or for functions that are not of bounded variation, the multifractal formalism for functions cannot be obtained as a consequence of the multifractal formalism for measures.

Our purpose in Part I of this paper is to prove Theorem 1.4. In section 2, we make explicit the relation between the size of the wavelet transform and the local regularity of the function. In section 3, we identify the quantities $S_q(l)$ or $Z(a, q)$ with some functional norms, thus proving the first point of Theorem 1.4. In section 4, we prove the upper estimate for the dimensions of singularities and the formula for the box dimension of the graph of F. In section 5, we study the wavelet-maxima method. In section 6, we construct counterexamples to the validity of the multifractal formalism in all generality.

The two parts of this paper can be read independently. Some results of this paper have been announced in [18], [19], and [20].

2. Regularity, singularities, and two-microlocalization. The results of Theorems 1.4 and 1.6 relate the pointwise behavior of a function to estimates on its wavelet transform. Our purpose in this section is to recall existing results on this topic and prove new ones concerning either negative exponents α or strong α-singularities. We first recall the basic properties of the wavelet transform.

Let ψ be in $C^{k+1}(\mathbb{R}^m)$, radial, with moments of order less than $k+1$ vanishing, and such that the derivatives of ψ of order less than $k+1$ have fast decay. The wavelet transform of F is defined by

$$C(a,b)(F) = \frac{1}{a^m}\int_{\mathbb{R}^m} F(t)\psi\left(\frac{t-b}{a}\right) dt, \tag{2.1}$$

and if $C(\psi) = \int |\hat{\psi}(\xi)|^2 d\xi/|\xi|$, F is recovered from its wavelet transform by

$$F(t) = C(\psi)\int_{a>0}\int C(a,b)(F)\psi\left(\frac{t-b}{a}\right) db\frac{da}{a^{m+1}}.$$

An intuitive idea is that a large wavelet coefficient means that the corresponding function locally has an oscillation at the corresponding scale of a corresponding amplitude. Although there does not seem to be a straightforward relationship between the two notions, Propositions 2.2 and 2.5 can be seen as a mathematical formulation of this idea. The following results can be found in [25] and [17]. Suppose that $s>0$.

- $F \in C^s(\mathbb{R}^m)$ if and only if

$$|C(a,b)(F)| \leq Ca^s. \tag{2.2}$$

(Recall that if $s=1$, the space $C^s(\mathbb{R}^m)$ must be replaced by the Zygmund class, which is composed of the continuous functions F such that $|F(x+h)+F(x-h)-2F(x)| \leq Ch$, or, more generally, if s is a positive integer, then it must be replaced by the corresponding indefinite integrals of the Zygmund class.)

- If $F \in C^s(x_0)$, then

$$|C(a,b)(F)| \leq Ca^s\left(1+\frac{|b-x_0|}{a}\right)^s. \tag{2.3}$$

- If (2.3) holds and if $F \in C^\varepsilon(\mathbb{R}^m)$ for an $\varepsilon>0$, there exists a polynomial P such that if $|x-x_0| \leq 1/2$,

$$|F(x)-P(x-x_0)| \leq C|x-x_0|^s \log\left(\frac{1}{|x-x_0|}\right). \tag{2.4}$$

Due partly to physical motivations (the study of the velocity of turbulent fluids, for instance), we do not want to consider only bounded functions, and thus we want to be able to consider points where F has a singularity (i.e., in a neighborhood in which it is unbounded). We first want to obtain results similar to (2.3) or (2.4) for singularities. A first problem is the definition of singularities that we should adopt.

The following definition is a straightforward generalization of (1.1) to negative exponents.

DEFINITION 2.1. *Suppose that $-m < s \leq 0$. F is $C^s(x_0)$ if*

$$|F(x)| \leq C|x - x_0|^s. \tag{2.5}$$

We have to make the assumption $-m < s \leq 0$ because if $s \leq m$, F might not be locally integrable and thus might not be a distribution. In that case, no computation on F (such as defining wavelet coefficients) would make sense. We will nonetheless see later how to define singularities of order less than $-m$.

We now relate (2.5) to conditions on the wavelet transform of F. We first check that if (2.5) holds, then

$$|C(a,b)(F)| \leq Ca^s \left(1 + \frac{|b - x_0|}{a}\right)^s; \tag{2.6}$$

First, suppose that ψ is supported in $B(0,1)$. Then

$$|C(a,b)| \leq \frac{C}{a^m} \int_{B(b,a)} |x - x_0|^s dx,$$

where $B(x,r)$ is the ball centered at x of radius r. If $|b - x_0| \geq 2a$ and $x \in B(b,a)$, then $|x - x_0| \sim |b - x_0|$ and $|C(a,b)|$ is bounded by $(C/a^m)4^m a^m |b - x_0|^s$. Otherwise, $|x - x_0| \sim a$ and the integral is bounded by $(C/a^m)4^m a^m a^s$, and hence we have (2.6). The general case holds because condition (2.6) does not depend on the particular wavelet chosen (see [21]).

Note that we will often use the notation $a \sim b$ for positive quantities, which will always mean that the quotient a/b is bounded from below and above by positive constants.

If (2.6) holds, one can easily check that it implies no regularity for F. In that case, of course, we refuse to make a minimal smoothness assumption like $F \in C^\varepsilon(\mathbb{R}^m)$, which was needed in a similar situation in order to get (2.4). Let us show intuitively how to obtain a converse estimate. Suppose that $\text{supp}\psi \subset B(0,1)$, (2.5) holds, and $|\nabla F(x)| \leq C|x - x_0|^{s-1}$; we further have $|C(a,b)| \leq Ca|b - x_0|^{s-1}$ for $|b - x_0| > a$.

Conversely, one can easily check that this last estimate together with (2.6) implies that $|F(x)| \leq C|x - x_0|^s$. We actually prove a slightly more general result.

PROPOSITION 2.2. *Let $-m < s \leq 0$. If $|F(x)| \leq C|x - x_0|^s$, then*

$$|C(a,b)(F)| \leq Ca^s \left(1 + \frac{|b - x_0|}{a}\right)^s.$$

Conversely, suppose that $\exists s' < s$ such that

$$|C(a,b)(F)| \leq Ca^s \left(1 + \frac{|b - x_0|}{a}\right)^{s'}. \tag{2.7}$$

Then $|F(x)| \leq C|x - x_0|^s$.

Proof. We already proved the first part. Suppose that (2.7) holds. Using the reconstruction formula for F,

$$|F(t)| \leq C \int \left[\int_{B(t,a)} |C(a,b)| db\right] \frac{da}{a^{m+1}}.$$

If $|t-x_0| \geq 2a$, $|b-x_0| \geq a$ and the right-hand side is bounded by

$$C \int_{a \leq \frac{|t-x_0|}{2}} a^{s-s'} |t-x_0|^{s'} \frac{a^m}{a^{m+1}} da \leq C|t-x_0|^s.$$

If $|t-x_0| \leq 2a$, $|b-x_0| \leq 4a$ and we get the bound

$$C \int_{a \geq \frac{|t-x_0|}{2}} a a^s a^m \frac{da}{a^{m+1}} \leq C|t-x_0|^s.$$

Hence Proposition 2.2 follows. □

Let us now recall the following definition of the two-microlocal spaces $C^{s,s'}(x_0)$ (see [17]):

$$F \in C^{s,s'}(x_0) \Longleftrightarrow |C(a,b)| \leq C a^s \left(1 + \frac{|b-x_0|}{a}\right)^{-s'}. \tag{2.8}$$

Proposition 2.2 generalizes to negative exponents the continuous embeddings

$$C^s(x_0) \hookrightarrow C^{s,-s'}(x_0) \quad \text{if } s' < s, \tag{2.9}$$

proved in [21], so that it also yields a justification of Definition 2.1 (and thus to the definition of strong α-singularities when $\alpha \leq 0$).

The problem of defining Hölder exponents for $s \leq -m$ is not straightforward. As mentioned before, we cannot consider only conditions such as $|F(x)| \leq C|x-x_0|^s$ since this does not imply that F is a distribution. The following definition has sometimes been proposed:

$$F \in C^s(x_0) \Longleftrightarrow (-\Delta)^{-\frac{[s]}{2}} F \in C^{s-[s]}(x_0). \tag{2.10}$$

There are two problems with this definition. The first is that it is not consistent with the definition for $s > 0$. Let us present an example. Consider $F(x) = x^{1/2}\cos(1/x)$; the integral of F is $O(x^{5/2})$ at the origin. Nonetheless, we would not consider F to be a $C^{3/2}$ function at the origin. Furthermore, this definition is also not consistent with the "natural" definition (2.5) when $-n < \alpha \leq 0$ for essentially the same reasons (we leave this verification to the reader). In order to go further, we interpret (2.10) as a two-microlocal condition. It implies $(-\Delta)^{-\frac{[s]}{2}} F \in C^{s-[s],-s+[s]}(x_0)$ so that $F \in C^{s,-s+[s]}(x_0)$. This condition is very far from $f \in C^{s,-s}$ which because of Proposition 2.2 should be "close" to the condition $F \in C^s(x_0)$. We show how to obtain a definition which is consistent with the definition for $s > -m$ and with the imbeddings in (2.9).

First, note that if s' is positive, $C^{s,s'}(x_0) \hookrightarrow C^s(\mathbb{R}^m)$, where by extension we define for a negative s

$$C^s(\mathbb{R}^m) = \dot{B}^{s,\infty}_{\infty} = \{F : \; |C(a,b)| \leq C a^s\}.$$

Thus the condition $F \in C^s(x_0)$, where s is negative, implies a global (negative) regularity for F. For $s \leq -m$, we will suppose that this regularity holds, which will guarantee that F is a distribution. In [11], Eyink proposed the definition $f \in C^{s,-s}(x_0)$. The advantage is that Proposition 4.1 can immediately be extended, which one uses with this definition of a pointwise Hölder exponent. The drawback is that this condition implies no pointwise regularity, even for positive s. Thus we adopt the following definition.

Definition 2.3. *Suppose that $s \leq -m$. F belongs to $C^s(x_0)$ if $F \in \dot{B}^{s,\infty}_{\infty}$ and if F restricted to $\mathbb{R}^m - \{x_0\}$ is a function that satisfies*

$$|F(x)| \leq C|x - x_0|^s.$$

Note that this definition is slightly redundant since any function defined on $\mathbb{R}^m - \{x_0\}$ is the restriction of a distribution (defined on $\mathbb{R}^m$) which belongs to $\dot{B}^{s,\infty}_{\infty}$.

If we define $\dot{C}^s(\mathbb{R}^m) = \dot{B}^{s,\infty}_{\infty}(\mathbb{R}^m)$, we have the surprising continuous embedding

$$\dot{C}^s(x_0) \hookrightarrow \dot{C}^s(\mathbb{R}^m),$$

which goes in the opposite direction than it would for positive s.

This definition coincides with Definition 2.1 when $-m < s \leq 0$ since in that case the function F itself is the corresponding distribution, so

$$|F(x)| \leq C|x - x_0|^s \implies F \in \dot{B}^{s,\infty}_{\infty}.$$

Suppose that $F \in C^s(x_0)$. If $|b - x_0| \geq 2a$, as in the case where $s > -m$, we get $|C(a,b)| \leq C|b - x_0|^s$. Since $|C(a,b)| \leq a^s$ by hypothesis, we see that $C^s(x_0) \hookrightarrow C^{s,-s}(x_0)$.

Proposition 2.4. *Using the previous definition of negative Hölder regularity, if $s \leq -m$, the following embeddings hold:*

$$\begin{cases} F \in C^s(x_0) \Rightarrow F \in C^{s,-s}(x_0), \\ F \in C^{s,-s'}(x_0) \quad \textit{for an } s' < s \Rightarrow F \in C^s(x_0). \end{cases} \tag{2.11}$$

The proof of the second implication is similar to the case where $s > -m$. It is interesting to check that some distributions which "should" belong to these generalized Hölder spaces satisfy these conditions. For instance, the distribution $p.p.(1/x)$ defined by

$$\left\langle p.p.\left(\frac{1}{x}\right) \mid \phi \right\rangle = \lim_{\epsilon \to 0} \int_{\mathbb{R} - [-\epsilon, \epsilon]} \frac{\phi(x)}{x} dx$$

is C^{-1} at 0 and $f.p.(1/x^2)$ defined by

$$\left\langle f.p.\left(\frac{1}{x^2}\right) \mid \phi \right\rangle = \lim_{\epsilon \to 0} \left(\int_{\mathbb{R} - [-\epsilon, \epsilon]} \frac{\phi(x)}{x^2} dx - \frac{2\phi(0)}{\epsilon} \right)$$

is C^{-2} at the origin. We leave these verifications as an exercise.

We now prove the following proposition, which relates the size of the wavelet transform to the existence of strong α-singularities when the wavelet used is compactly supported.

Proposition 2.5. *Suppose that F is $C^\alpha(x_0)$ and that x_0 is a wavelet α-singularity of F. Then x_0 is a strong α-singularity of F.*

For the sake of simplicity, we restrict our focus to the case where $0 < \alpha < 1$. Suppose that F is $C^\alpha(x_0)$ and that x_0 is not a strong α-singularity of F. Let $\epsilon > 0$

be fixed and $a > 0$ be such that (1.3) does not hold for any (a, b) in the cone over x_0. For any x in the ball $B(x_0, Ca)$ (except on an exceptional set E_a of measure at most ϵa^m), we have

$$|F(x) - \bar{F}| \leq \epsilon a^{\alpha},$$

where $\bar{F}$ is the mean value of F in the ball $B(x_0, Ca)$. Also, if $x \in E_a$,

$$|F(x) - \bar{F}| \leq |x - x_0|^{\alpha}.$$

If the support of the wavelet $\psi((x - b)/a)$ is included in $B(x_0, Ca)$,

$$|C(a,b)| = \frac{1}{a^m} \left| \int F(x) \psi\left(\frac{x-b}{a}\right) dx \right| \leq \frac{1}{a^m} \int_{B(x_0, Ca)} |F - \bar{F}|,$$

the integral on E_a is bounded by $a^{\alpha} \epsilon a^m$ and outside E_a by $\epsilon a^{\alpha} a^m$, so $|C(a, b)| \leq 2C\epsilon a^{\alpha}$ and (1.3) does not hold. Hence we have a contradiction, and thus Proposition 2.5 holds.

The condition that F is $C^{\alpha}(x_0)$ is necessary in Proposition 2.5, as shown by the following counterexample. Suppose that ψ (perhaps after a translation) is compactly supported in an interval of the form $[2^l, 2^{l+1}]$, and suppose that the $2^{j/2}\psi(2^j x - k)$'s form an orthonormal wavelet basis of $L^2(\mathbb{R})$ (see [8] for such functions). Let I be an interval such that $\psi(x) \geq C > 0$ on I. Define $F(x) = \sum_j 2^{-(\alpha-1)j} 1_{A_j}(x)$, where $A_j = 2^{-j} I_j$ and I_j is a subinterval of I of length 2^{-j}. Then clearly $2^{-j} \int F(x)\psi(2^j x) dx \geq C 2^{-\alpha l j}$ but F has no strong singularity at 0 (but is only $C^{\alpha-1}(0)$).

3. Some functional norm estimates. We first show the link between quantities such as $S_p(l)$ or $\tilde{Z}(a, q)$ and Sobolev or Besov-type norms. We recall a few definitions and characterizations.

Suppose that $s \in \mathbb{R}$ and $p, q > 0$. A function F belongs to the homogeneous Besov space $B_p^{s,q}$ if

$$\int_{a>0} \left[\int |C(a,b)|^p db \right]^{q/p} \frac{da}{a^{sq+1}} < +\infty \tag{3.1}$$

(which follows directly from [25]).

Since $\eta(p)$ is the infimum of all numbers τ verifying, for a small enough,

$$\tilde{Z}(a,p) \left(= \int |C(a,b)|^p db \right) \leq C a^{\tau},$$

we see that if $p > 0$,

$$\eta(p) = \sup\{\tau : F \in B_p^{\tau/p, \infty}\}. \tag{3.2}$$

A similar characterization exists for the function $\zeta(p)$. The spaces $H^{s,p}$ introduced by Nikol'skii (see [1] or [27]) are defined as follows.

Let $s \geq 0$. If s is not an integer, $s = m + \sigma$ with m integer and $0 < \sigma < 1$. Let $p \geq 1$, $F \in H^{s,p}$ if $F \in L^p$ and for any multiindex α such that $|\alpha| = m$,

$$\int \frac{|\partial^{\alpha} F(x+h) - \partial^{\alpha} F(x)|^p}{|h|^{\sigma p}} dx \leq C. \tag{3.3}$$

Recall that $\zeta(p)$ is the lim sup of the numbers ξ such that

$$S_p(h)\left(=\int |F(x+h)-F(x)|^p dx\right) \leq Ch^{\xi(p)}$$

for h small enough. Thus if $p \geq 1$, $\zeta(p) = \sup\{s : F \in H^{s/p,p}\}$.

Of course, we see here that the formula in the structure function method must be modified as follows in order to be consistent with (3.3): If $\zeta(p)$ is less than 1, the formula is all right; if it is equal to 1, one should use the same formula but with the gradient of F; and so on until $\zeta(p)$ falls between two integers. (Note that this procedure is obviously difficult to handle numerically if $\zeta(p)$ is large.)

The following embeddings hold if $p \geq 1$:

$$\forall \epsilon > 0, \quad H^{s+\epsilon,p} \hookrightarrow B_p^{s,\infty} \hookrightarrow H^{s-\epsilon,p} \tag{3.4}$$

(because (3.4) holds between $H^{s,p}$ and $W^{s,p}$ spaces (see [1]), between $W^{s,p}$ and $L^{p,s}$ spaces (see [1]), and between $L^{p,s}$ and $B^{s,\infty}$ spaces (see [21] or [25])). Thus, if $p > 1$, $\zeta(p) = \eta(p)$ and the function η can be defined by

$$\eta(p) = \sup\{s : F \in B_p^{s/p,p}\} = \sup\{s : F \in L^{p,s/p}\} \tag{3.5}$$

(where $L^{p,s}$ is defined for $s > 0$ by $f \in L^{p,s} \Leftrightarrow f \in L^p$ and $(-\Delta)^{s/2} f \in L^p$), and if $0 < p \leq 1$, it can be defined by the first equality only, so the last characterization of $\eta(p)$ in (3.5) is again a straightforward consequence of Sobolev-type embeddings.

PROPOSITION 3.1. *The following characterizations hold:*

$$\forall p > 0, \quad \eta(p) = \sup\{s : F \in B_p^{s/p,\infty}\},$$
$$\forall p > 1, \quad \eta(p) = \zeta(p) = \sup\{s : F \in H^{s/p,p}\} = \sup\{s : F \in L^{s/p,p}\}.$$

Remark. The number $\eta(2)$ can be interpreted as follows:

$$\eta(2) = \sup\left\{s : \int |\hat{F}(\xi)|^2 (1+|\xi|^2)^{s/2} d\xi \leq C\right\}.$$

This holds because $B_2^{s,2} = L^{2,s}$, and $\forall q, q', q''$, $B_p^{s+\epsilon,q} \subset B_p^{s,q'} \subset B_p^{s-\epsilon,q''}$, so

$$\begin{aligned}\eta(2) &= \sup\{s : F \in B_2^{s,\infty}\} = \sup\{s : F \in B_2^{s,2}\} \\ &= \sup\{s : F \in L^{2,s/2}\} = \sup\left\{s : \int |\hat{F}(\xi)|^2 (1+|\xi|^2)^{s/2} d\xi \leq C\right\}.\end{aligned}$$

Note that this result differs from [3], where the interpretation given for $\eta(2)$ is $|\hat{F}(\xi)|^2 \sim |\xi|^{-\eta(2)-2}$. Nonetheless, the interpretation given in [3] is correct provided that such a scaling holds. An interpretation of $\eta(1)$ of very different nature will be given in section 4.

We will show in section 6 that "any" function $d(\alpha)$ can be a Hölder spectrum. It is interesting to notice that this is not the case with the function $\eta(p)$, which because of the Sobolev imbeddings between $L^{p,s}$ spaces cannot be arbitrary. Since $L^{p,s} \subset L^{t,q}$ if $t \leq s$ and $q = mp/(m-(s-t)p)$ (see [1]), if $q \geq p$,

$$\frac{\eta(q)-\eta(p)}{q-p} \geq \frac{\eta(p)-m}{p}. \tag{3.6}$$

In particular, we see that $\eta'(p) \geq \eta(p) - m/p$. Conversely, it is easy to check that any function $\eta(p)$ that satisfies (3.6) can be associated with a function F so that (3.6) characterizes all possible functions $\eta(p)$.

4. Upper bounds for dimensions of spectrums. A first problem that we meet is that of which mathematical definition of "dimension" we should use. The physical literature is often unclear about this point, sometimes using the term *Hausdorff dimension* but computing it using coverings by boxes of the same size. Of course, a given set of points (a potential "set of Hölder singularities" of our function) can have very different dimensions depending on the definition considered. We will see that the "good definition" depends on the kind of singularities that we look for. For Hölder singularities, we will get bounds on Hausdorff dimensions, and for strong α-singularities, we will get bounds on packing dimensions. An important difference between the two settings is that in the first we necessarily have to suppose some minimal uniform regularity for F, which is not required in the second. We first recall the definition of the Hausdorff dimension and Hausdorff measure.

Let $A \subset \mathbb{R}^n$ and R_ε be the set of all coverings of A by sets of diameter at most ε. Let

$$M(\varepsilon, d) = \inf_{r \in R_\varepsilon} \sum_{A_i \in r} (\operatorname{diam} A_i)^d.$$

Then by definition,

$$d - \operatorname{Mes}(A) = \limsup_{\varepsilon \to 0} \; M(\varepsilon, d)$$

is the d-dimensional Hausdorff measure. The *Hausdorff dimension* of A is

$$D = \inf\{d : d - \operatorname{Mes}(A) = 0\} = \sup\;\{d : d - \operatorname{Mes}(A) = +\infty\}.$$

If the coverings are done using only balls or only dyadic cubes, we obtain an equivalent quantity for the d-measure, and thus D is not changed.

Proposition 4.1. *Let $s - m/p > 0$ and $p > 0$. If $F \in B_p^{s,\infty}$, $d(\alpha) \le m - (s-\alpha)p$. Thus if $\eta(p)$ satisfies $\eta(p) > m \;\forall p$, $d(\alpha) \le \inf_p(m - \eta(p) + \alpha p)$.*

This proposition is reminiscent of [5], where Brown, Michon, and Peyrière proved similar results for measures (in dimension 1). If $s \le m/p$, a function in $L^{p,s}$ or $B_p^{s,\infty}$ can be infinite on a dense set and thus smooth at no point (see [21]), so that no such result can hold if we do not make the assumption $s - m/p > 0$.

Proof of Proposition 4.1. we use a slight modification of the two-microlocal space, for convenience. We thus define

$$(4.1)\quad F \in C_p^{s,s'}(x_0) \quad \text{if and only if } |C_{j,k}| \le C 2^{-(\frac{m}{2}+s)j} j^{2/p} (1 + |2^j x - k|)^{-s'}.$$

We will prove that if $F \in B_p^{s,\infty}$, then $d > 0$. Outside a set of d-measure 0, $F \in C_p^{s-m-d/p,-d/p}(x)$. Thus if $0 < s - m/p < \alpha < s$, the set $\{x : \;\; F \notin C^\alpha(x)\}$ has Hausdorff dimension at most $m - (s-\alpha)p$, and Proposition 4.1 follows.

Let $F \in B_p^{s,\infty}$. Then

$$(4.2)\qquad \forall j, \quad \sum_k |C_{j,k}|^p \, 2^{(ps + \frac{mp}{2} - m)j} \le C.$$

Let d be such that $0 < d \le n$ and $B_{j,k}$ be the ball centered on $k2^{-j}$ and of size

$$\operatorname{diam}(B_{j,k}) \;=\; |C_{j,k}|^{p/d} 2^{\alpha j} j^{-2/d},$$

where α is such that

$$-d\alpha + ps + \frac{mp}{2} - m \;=\; 0.$$

Then (4.2) can be rewritten as

$$\forall j, \quad \sum_k (\operatorname{diam} B_{j,k})^d \le \frac{c}{j^2}. \tag{4.3}$$

Let $A_j = \bigcup_k B_{j,k}$. Equation (4.3) implies that the d-measure of $A = \limsup A_j$ is 0. If $x \notin \limsup A_j$, $\exists\, j_0,\ \forall j \ge 0,\ \forall k,\ x \notin B_{j,k}$ so that

$$|x - k2^{-j}| \ge C|C_{j,k}|^{p/d}\, 2^{\alpha j} j^{-2/d}.$$

Hence

$$\forall j \ge j_0, \quad |C_{j,k}| \le C2^{-(m/2+s-m/p)j}|x - k2^{-j}|^{d/p} j^{2/p}$$

and thus $F \in C_p^{s-m-d/p\,,-d/p}(x)$ (because (4.1) automatically holds for $j \le j_0$). Hence Proposition 4.1 follows.

One can wonder if similar bounds (or equalities) hold for dimensions of strong α-singularities. This problem is important for the following reasons. Recall that the multifractal formalism was introduced for the study of turbulence. In [6], Caffarelli, Kohn, and Nirenberg obtained a bound on the dimension of (possible) singularities in Navier–Stokes equations that is actually a bound on the packing dimension of "strong α-singularities" following the definition that we gave (with $\alpha = 0$).

Another reason to obtain bounds for dimensions of strong singularities is that when global regularity conditions (which imply that F is continuous) no longer hold, no result such as Proposition 4.1 can be proved. Even in the strict framework of self-similar functions, we will see in Part II that no such bounds exist. Since for applications we clearly want to be able to consider unbounded functions (for instance, the velocity of a turbulent fluid may be unbounded), it is important to obtain some positive results in that case.

Our purpose is to prove that if F belongs to $W^{s,p}$, given $\alpha < s$, the set of points x where F has a strong α-singularity has a small packing dimension. We first recall the definition of the packing dimension of a subset of $\mathbb{R}^m$ (see [12]).

Let $J > 0$ and Λ_J be the set of dyadic cubes of size 2^{-J} which contain a point of E. Define

$$m_d(E) \;=\; \lim_{J\to+\infty} \sum_{\lambda\in\Lambda_J} 2^{-dJ} \;=\; \Lambda_j^\sharp\, 2^{-dJ}$$

(where $\Lambda_j^\sharp$ denotes the cardinality of Λ_j) and

$$\operatorname{mes}_d(E) \;=\; \inf_{E\subset\cup E_n} \sum_n m_d\,(E_n).$$

The *box dimension* of E is the value of d for which $m_d\,(E)$ falls from $+\infty$ to 0. This dimension is also called the potential dimension by some physicists. It is the only one that is numerically easy to compute because it does not involve optimal coverings.

The *packing dimension* of E is the value of d for which $\mathrm{mes}_d(E)$ falls from $+\infty$ to 0. It is clearly no larger than the box dimension.

Proposition 4.2. *Let $F \in W^{s,p}(\mathbb{R}^m)$ and α be such that $-m < \alpha \le 1$. The packing dimension of the strong α-singularities of F is bounded by $m-(s-\alpha)p$ so that if $D(\alpha)$ is the packing dimension of strong α-singularities and $-m < \alpha \le 1$, then*

$$D(\alpha) \le \inf_p (m - \eta(p) + \alpha p). \tag{4.4}$$

Such a result is in many cases more satisfactory than Proposition 4.1 since we do not have to make the assumption of a minimal Hölder regularity of F. Actually, the numerical estimation of the upper bound for $D(\alpha)$ when $\alpha = 0$ using (4.4) could be a way to check whether a stronger result than the one obtained by Caffarelli, Kohn, and Nirenberg in [6] holds.

Of course, a way to avoid the problem of unbounded functions could be to consider indefinite integrals or perhaps iterated indefinite integrals of the velocity, but such quantities would have no direct physical interpretation.

We first describe the functional setting that we use. We will give bounds on the packing dimension of strong α-singularities in the Sobolev spaces $W^{s,p}$. Recall that (see [1])

(4.5)

$$\text{if } \ 0 < s < 1, \quad f \in W^{s,p} \ \Leftrightarrow \ f \in L^p \quad \text{and} \quad \int\!\!\int \frac{|f(x+t) - f(x)|^p}{|t|^{m+sp}}\, dx\, dt \le +\infty.$$

For $s \ge 1$, these spaces can be defined as follows. First, if $0 < s < 2$, they can be defined by replacing $|f(x+t) - f(x)|$ by $|f(x+t) + f(x-t) - 2f(x)|$ in (4.5), and if $\alpha \ge 2$, $f \in W^{s,p} \Leftrightarrow f \in L^p$ and $\forall i = 1, \dots, n$, $\partial f/\partial x_i \in W^{s-1,p}$ (see [1]).

The fact that these spaces are defined by a condition on the L^p-modulus of continuity $\omega_p(t) = \|f(\cdot + t) - f(\cdot)\|_p$ will yield an easy direct estimate on the packing dimension of the strong α-singularities. (The intuitive idea is that if x_0 is such a singularity, the contribution for x close to x_0 to the integral $\int_{x\in\mathbb{R}^m} |f(x+t) - f(x)|^p\, dx$ is large.) The spaces $L^{p,s}$ and $W^{s,p}$ are closely related since (see [1]) $W^{s,2} = L^{2,s}$ and $L^{p,s} \subset W^{s,p'}$ if $p > p'$ and $W^{s,p} \subset L^{p',s}$ if $p > p'$. Thus the bound of $D(\alpha)$ given by Theorem 1.4 is a consequence of Proposition 4.2 which we now prove. Define $E_{l,m}$ as the set of points x_0 such that (1.2) holds with

$$2^{-l} \le C < 2 \cdot 2^{-l} \quad \text{and} \quad 2^{-n} \le C' < 2 \cdot 2^{-n}.$$

The set of strong α-singularities of F is $\bigcup_{l,n} E_{l,n}$. Let l and n be fixed. Let Λ_j be the set of dyadic cubes of size 2^{-j} such that $\Lambda_j \cap E_{l,n} \ne \emptyset$. If $\lambda \in \Lambda_j$, there exist $A_j, B_j \subset 3\lambda$ such that (1.2) holds (where 3λ is the cube that has the same center as λ and is three times larger). We restrict the integral (4.5) to $x \in 3\lambda$, $x + t \in 3\lambda$. The integral on this set is thus bounded from below by

$$2^{-2l} 2^{-2mj} \frac{(2^{-n} 2^{-\alpha j})^p}{(2^{-j})^{m+sp}}.$$

If we sum up for all $\lambda \in \Lambda_j$, each integral is taken at most 4^m times. Thus

$$\Lambda_j^\sharp \, 2^{-2l} \, 2^{-np} \, 2^{-j[\alpha p + m - sp]} \le 4^m \|f\|^p_{W^{s,p}}$$

and thus if $d > \alpha p + m - sp$, $d - \text{mes}(E_{l,m}) = 0$ so that $d - \text{mes}(\bigcup_{l,n} E_{l,n}) = 0$. Hence Proposition 4.2 follows.

We now check that if F is a function of one real variable, the box dimension of the graph of F is exactly $2 - \eta(1)$ if $\eta(1)$ is between 0 and 1. This is a straightforward consequence of the following result (see [10] or [13]).

Proposition 4.3. *Suppose* $0 < \gamma < 1$ *and* $F : [0, 1] \to \mathbb{R}$ *is continuous. Then the box dimension of the graph of* F *is exactly* $2 - \gamma$ *if and only if*

$$F \in \bigcap_{\alpha < \gamma} B_1^{\alpha,\infty} \Big\backslash \bigcup_{\beta > \gamma} B_1^{\beta,\infty}.$$

Thus the result holds because $\eta(p) = \sup\{s : F \in B_p^{s/p,\infty}\}$.

5. The wavelet-maxima method. Our purpose in this section is to show that the wavelet-maxima method can yield a function $\theta(q)$ which is much smaller than $\eta(q) - m$ so that in general the multifractal formalism cannot hold using this method. Via our counterexamples, we will show how to slightly modify its definition so that $\theta(q) = \eta(q) - m$. Our specific study of the wavelet-maxima method is justified by its numerical importance. Arneodo, Bacry, and Muzy compared the three numerical methods in cases where the Hölder spectrum is known analytically (self-similar functions, Riemann's function), and they clearly showed (in a personal communication) that the wavelet-maxima method is the most accurate.

The reason why the wavelet-maxima method may fail is easy to understand intuitively if we relate it to the wavelet-transform integral method. The two quantities $\int_{\mathbb{R}^m} |C(a, b)|^q db$ and $a \sum_{\ell \in \mathcal{L}(a)} \sup_{(b,a') \in \ell} |C(a', b)|^q$ have the same order of magnitude if the spacing between the maxima is approximately a since then the second term is a Riemann sum of the first term. Thus the counterexamples that we will construct will have maxima with spacing much smaller than a, and if we slightly modify the wavelet-maxima method by imposing the restriction that we select only one maximum (or, say, C maxima) in an interval of length a, then the multifractal formalism will hold.

In order to give some insight into the pitfalls of the wavelet-maxima method, we begin by describing an example where the maxima accumulate in certain regions. Not surprisingly, this example involves chirps.

Lemma 5.1. *Suppose that* ψ *is compactly supported on* $[0, l]$, *has a vanishing integral and* m *first vanishing moments, and satisfies*

$$\exists \epsilon > 0 \quad \psi(x) = x^m \quad \forall x \in [0, \epsilon].$$

(This is the case, for instance, if ψ *is a spline.) There exists a function* F *that is compactly supported and arbitrarily smooth and a sequence* $a_n \to 0$ *such that for all values of* n, *the wavelet transform* $C(a_n, b)$ *has infinite maxima.*

We first construct F such that this property holds for a small interval of values of the dilation parameter a. The general case will be obtained by a superposition argument. Let

$$F(x) = x^k \sin\left(\frac{1}{x^l}\right) \phi(x),$$

where ϕ is C^∞ except at the origin and supported on $[0, 1]$, $\phi(x) = 1$ $\forall x \in [0, 1/2]$, and ϕ is such that the integral and the first m moments of F vanish. After dilating

ψ, we can suppose that it is equal to x^m on the interval $[0,1]$. Then if $x \leq 1/4$ and $a \in [1/2,1]$,

$$\begin{aligned}\frac{1}{a}\int F(t)\psi\left(\frac{x-t}{a}\right)dt &= \frac{1}{a}\int_x^1 F(t)\left(\frac{x-t}{a}\right)^m dt\\ &= -\frac{1}{a}\int_0^x F(t)\left(\frac{x-t}{a}\right)^m dt = -\frac{1}{a}\int_0^x t^k\left(\frac{x-t}{a}\right)^m \sin(t^{-l})dt\end{aligned}$$

Integrating m times by parts, we obtain either

$$\int F(t)\psi(x-t)dt = a^{-m-1}x^{k+m(l+1)}\sin\frac{1}{x^l} + o(x^{k+m(l+1)})$$

or

$$\int F(t)\psi(x-t)dt = a^{-m-1}x^{k+m(l+1)}\cos\frac{1}{x^l} + o(x^{k+m(l+1)})$$

depending on the parity of m. In all cases, the wavelet transform of F has for $a \in [1/2,1]$ an infinity of lines of maxima. The general case is obtained by considering the function

$$G(x) = \sum_{j=0}^{\infty} 2^{-mj}F(2^j(x-l)),$$

where l is larger than the size of the support of ψ.

This example also shows that one should be careful when using the wavelet-maxima method since the superposition of a small smooth function can completely perturbate the lines of maxima.

We now show that the two functions $\theta(q)$ and $\eta(q)-1$ can differ dramatically so that even in cases where the multifractal formalism holds when using the wavelet integral method, it may prove wrong when using the wavelet-maxima method. To this end, we will construct a smooth function F (so that $\eta(p)$ will take the maximal value that is compatible with the smoothness of the wavelet) such that $\theta(q) = -\infty\ \forall q$. This example will use a wavelet with one vanishing moment. However, we will show how to modify it in order to deal with wavelets with a given number of vanishing moments. We will also show in Part II that F can be a self-similar function (which will provide a case where the multifractal formalism holds using the wavelet integral method and does not hold using the wavelet-maxima method).

Proposition 5.2. *Let ψ be even and compactly supported* (*say on* $[-1,1]$) *and satisfy*

$$\int \psi(x)dx = 0 \quad \textit{and} \quad \int x\psi(x)dx = 1.$$

There exists a C^∞ compactly supported function F such that $\theta(q) = -\infty\ \forall q > 0$.

Proof. The idea of the proof is to construct a function g whose wavelet transform is equal to, say, 1 on an interval and to perturbate it by adding another function whose wavelet transform is extremely small but oscillates extremely fast, thus creating a huge number of new maxima which take values close to 1.

Let g be a C^∞ odd function supported by $[-3,3]$ such that $g(x)=1$ on $[1,2]$. Let

$$h_j(x) = 2^{-j^2}g(2^{j+4}(x-8)) + 2^{-j^4}\sin(2^{j^3}\pi x)\phi(2^jx),$$

where ϕ is a C^∞ function supported on $[1/2, 1]$ that verifies $\phi(x+3/4) = \phi(3/4-x)$ and $\forall x \in [9/16, 15/16]$, $\phi(x) = 1$. Let F be the indefinite integral of $\sum_{j\geq 0} h_j(x)$. Since h_j has a vanishing integral, F is C^∞ and compactly supported. Let G be the indefinite integral of g. F and the series $\sum_{j\geq 0} 2^{-j^2}2^{-j-4}G(2^{j+4}(x-8))$ will have the same function η. (Here the calculation will yield $\eta(p) = p$ because this series is a C^∞ function and the wavelet used will have only one vanishing moment.)

Note that

$$\frac{1}{a}\int \sin(\omega x)\psi\left(\frac{x-b}{a}\right)dx \sin(\omega b)\hat{\psi}(\omega a).$$

(Here $\omega = 2^{j^3}$.) For a given value of j, we choose a in the interval $[1/100.2^{-j}, 1/10.2^{-j}]$ such that $\hat{\psi}(\omega a)$ does not vanish (which is possible since $\hat{\psi}(\omega a)$ is an analytic function of a).

Integrating by parts, one checks that on an interval of length at least $2^{-j-4}5/16$, the wavelet transform of $2^{-j^2}2^{-j-4}G(2^{j+4}(x-8))$ takes a constant value equal to 2^{-j^2}. Thus on the same interval, the wavelet transform of F is $2^{-j^2} + 2^{-j^4}\sin(2^{j^3}b)\hat{\psi}(2^{j^3}a)$. Thus it has about $2^{-j-4}2^{j^3}$ maxima, and

$$\sum_{\max}|C(a,b)|^q \sim 2^{-j-4}2^{j^3}2^{-j^2q}.$$

Since j can be chosen arbitrarily large, the result is proved. □

Note that we could have chosen a wavelet with a given number of vanishing moments. In that case, we would have integrated g not once but the corresponding number of times. The important fact is that the wavelet transform of G should locally be constant. The reader will also easily check that we could have imposed a given function η for F.

6. Counterexamples to the multifractal formalism. We define $\mathcal{C}$ as the class of functions that can be written as the supremum of a countable set of functions of the form $c1_{[a,b]}(x)$ (where we can have $a = b$). Thus Riemann-integrable functions belong to $\mathcal{C}$, but so do, for instance, the indicatrix function of the rationals (but not the indicatrix function of the irrationals).

Proposition 6.1. *Let $d(s) :]0, +\infty[\to [0, m]$ be a function in $\mathcal{C}$. There exist two continuous functions G_1 and $G_2 : \mathbb{R}^m \to \mathbb{R}$ that share the same function $\eta(q)$ such that $d(s)$ is the Hölder spectrum of G_1 while G_2 is C^∞ except at the origin, so its spectrum vanishes everywhere.*

We construct these functions when the space dimension is $m = 1$. The generalization to the multidimensional case is straightforward.

We first construct G_1 when $d(s) = cs1_{a,b}(s)$, where $0 < a \leq b < \infty$ and $cb \leq 1$. We will actually use three other parameters α, β, and γ, where $a = \gamma$, $b = \beta\gamma$, and $c = 1/(\alpha\beta\gamma)$ so that $\gamma > 0$, $\beta \geq 1$, and $\alpha \geq 1$. We thus define $G_1 = F^{(\alpha,\beta,\gamma)}$. The general case will be obtained using a simple "superposition" procedure of the $F^{(\alpha,\beta,\gamma)}$.

We will explicitly construct G_1 by defining its coefficients on an orthonormal wavelet basis. The function G_2 will then be obtained by just moving at each scale the location of the nonvanishing wavelet coefficients of G_1. We use an orthonormal wavelet basis in the Schwartz class (see [25]), and the functions

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^jx - k), \quad j, k \in \mathbf{Z},$$

are an orthonormal basis of $L^2(\mathbb{R})$. Sometimes we will index the wavelets $\psi_{j,k}$ or the wavelet coefficients $C_{j,k}$ $(= \int F\psi_{j,k})$ by the dyadic intervals $\lambda = [k2^{-j}, (k+1)k2^{-j}]$.

Let Λ be the collection of all dyadic intervals of length at most 1. We will construct a subcollection $\Lambda(\alpha,\beta) \subset \Lambda$ and consider the following "lacunary" wavelet series:

$$F(x) = \sum_{\lambda\in\Lambda(\alpha,\beta)} 2^{-(\gamma+1/2)j}\psi_\lambda(x). \tag{6.1}$$

The construction of $\Lambda(\alpha,\beta)$ is performed as follows. Define $\Lambda(\alpha,\beta) = \bigcup_{m\geq 1} \Lambda_m^{(\alpha,\beta)}$, where $\Lambda_m^{(\alpha,\beta)}$ is the set of intervals λ or couples (j,k) such that $j = [\alpha\beta m]$ and

$$2^{-j}k = \epsilon_1 l_1 + \cdots + \epsilon_m l_m \in F_m, \quad \epsilon_1,\dots,\epsilon_m \in \{0,1\}, \quad l_n = 2^{-[\alpha n]},$$

$[x]$ is the entire part of x and thus $k = 2^{[\alpha\beta m]}(\pm l_1 \pm \cdots \pm l_m)$ is an integer since $[\alpha\beta m] \geq [\alpha n]$.

Proposition 6.2. *The function F defined by* (6.1) *belongs to the global Hölder space $C^\gamma(\mathbb{R})$ so that if $s<\gamma$, the set $E^{(s)}$ of points x_0 where $f\in\Gamma^s(x_0)$ is empty. If $\gamma\leq s\leq\beta\gamma$, the Hausdorff dimension of $E^{(s)}$ is $s/\alpha\beta\gamma$. If $s>\beta\gamma$, $E^{(s)}$ is empty.*

The characterization of the space C^γ on the wavelet coefficients is

$$|C_{j,k}| \leq C2^{-(\gamma+1/2)j}$$

(a simple rewriting of (2.2) in the orthonormal basis setting). Thus F belongs to $C^\gamma(\mathbb{R})$ and the spectrum of F vanishes for $s<\gamma$.

Lemma 6.3. *A point x_0 belongs to $E^{(s)}$ if and only if*

$$\operatorname{dist}(x_0, F_m) = \eta_m 2^{-\frac{\alpha\beta\gamma}{s}m}, \quad (F_m = \{\pm l_1 \pm l_2 \pm \cdots \pm l_m\}) \tag{6.2}$$

with

$$\liminf_{m\to\infty} \eta_m 2^{-m\epsilon} = 0 \quad \textit{for any } \epsilon > 0 \tag{6.3}$$

and

$$\liminf_{m\to\infty} \eta_m 2^{m\epsilon} = +\infty \quad \textit{for any } \epsilon > 0. \tag{6.4}$$

Proof. If F is $C^s(x_0)$, the rewriting of (2.3) yields

$$|C_{j,k}| \leq C2^{-(s+1/2)j}(1+|2^jx_0-k|)^s. \tag{6.5}$$

Conversely, from (2.3), we deduce that if (6.5) holds and if F is $C^\epsilon(\mathbb{R})$ for an $\epsilon>0$, then there exists a polynomial P such that

$$|F(x)-P(x-x_0)| \leq C|x-x_0|^s \log\left(\frac{1}{|x-x_0|}\right).$$

Thus we see that F is $C^{s-\epsilon}(x_0)$ $\forall\epsilon>0$ if $\forall\epsilon>0$, $\forall\lambda\in\Lambda(\alpha,\beta)$,

$$2^{-(\gamma+1/2)j} \leq C2^{-(s-\epsilon+1/2)j}(1+|2^jx_0-k|)^{s-\epsilon} = C2^{-j/2}(2^{-j}+\operatorname{dist}(x_0,\lambda))^{s-\epsilon}.$$

Conversely, if $\exists\lambda\in\Lambda(\alpha,\beta)$ corresponding to arbitrary large values of j such that

$$2^{-(\gamma+1/2)j} \geq C2^{-j/2}(2^{-j}+\operatorname{dist}(x_0,\lambda))^{s+\epsilon},$$

F does not belong to $C^{s+\epsilon}(x_0)$ $\forall\epsilon > 0$. These two conditions can be written as

$$\limsup_{\lambda\in\Lambda(\alpha,\beta)} 2^{-\gamma j}(2^{-j} + \operatorname{dist}(x_0,\lambda))^{-s-\epsilon} = +\infty \quad \text{for any } \epsilon > 0 \tag{6.6}$$

and

$$\limsup_{\lambda\in\Lambda(\alpha,\beta)} 2^{-\gamma j}(2^{-j} + \operatorname{dist}(x_0,\lambda))^{-s+\epsilon} < \infty \quad \text{for any } \epsilon > 0. \tag{6.7}$$

Condition (6.6) can also be written as

$$2^{-j} + \operatorname{dist}(x_0,\lambda) = \eta(\lambda)2^{-\frac{\gamma}{s+\epsilon}j},$$

where $\liminf \eta(\lambda) = 0$. Since $s \geq \gamma$, $2^{-j} = o(2^{-\frac{\gamma}{s+\epsilon}j})$ and the only condition to be checked is

$$\operatorname{dist}(x_0,\lambda) = \eta(\lambda)2^{-\frac{\gamma}{s+\epsilon}j}.$$

Since $\lambda \in \Lambda_m^{(\lambda,\beta)}$, this condition is equivalent to (6.3).

The same proof shows that (6.7) becomes (6.4). Hence we have Lemma 6.3.

We now define a compact K_α and sets $E_{\alpha,\delta}$, K_α will be composed of the limit points of the F_m, and the $E_{\alpha,\delta}$'s will be subsets of K_α.

Let K_α be the compact set of the sums $\sum_1^\infty \epsilon_j l_j$, where $\epsilon_j = \pm 1$. Another equivalent definition is

$$K_\alpha = \bigcap_1^\infty (F_m + [-\lambda_m, \lambda_m]),$$

where

$$\lambda_m = l_{m+1} + l_{m+2} + \cdots.$$

Note that the sets $G_m = F_m + [-\lambda_m, \lambda_m]$ form a decreasing sequence of compact sets.

Let $G_m^{(\beta)}$ be defined by $G_m^{(\beta)} = F_m + [-\lambda_m^\beta, \lambda_m^\beta]$ and let $E_{\alpha,\beta}$ be the set of points that belong to infinite $G_m^{(\beta)}$'s. Since $\beta \geq 1$, $G_m^{(\beta)} \subset G_m$ so that $E_{\alpha,\beta} \subset K_\alpha$ and, of course, $E_{\alpha,\beta} = K_\alpha$ if $\beta = 1$.

The idea of the construction that we made is as follows. We have placed "large" wavelet coefficients on F_m so that on these sets the function F is exactly Γ^γ, but at points which are at a certain distance on F_m (measured by their belonging to certain $G_m^{(\beta)}$'s), these "large" wavelet coefficients create "weaker" singularities (corresponding to an exponent larger than γ).

Lemma 6.4. *If $\gamma \leq s < \beta\gamma$, then* (6.3) *is equivalent to $x \in \bigcap_{\delta<\frac{\beta\gamma}{s}} E_{\alpha,\delta}$, while if $s \geq \beta\gamma$, it is equivalent to $x \in K_\alpha$. Condition* (6.4) *is equivalent to $x \notin \bigcup_{\delta>\frac{\beta\gamma}{s}} E_{\alpha,\delta}$, while if $s > \beta\gamma$, it is equivalent to $x \notin K_\alpha$.*

Proof. If (6.3) holds and if $\delta < \beta\gamma/s$, let us check that $x \in E_{\alpha,\delta}$. To this end, we choose $\epsilon > 0$ such that $\delta < \beta\gamma/s - \epsilon$. Then

$$\operatorname{dist}(x_0, F_m) = \eta_m 2^{-\frac{\alpha\beta\gamma}{s}m} = o(2^{-(\frac{\alpha\beta\gamma}{s}-\epsilon)m})$$

so that $\operatorname{dist}(x_0, F_m) = o(l_m^\delta) \leq \lambda_m^\delta$ (because $\lambda_m \sim l_m$) for infinite values of m. Thus $x \in E_{\alpha,\delta}$.

Conversely, if $x \in E_{\alpha,\delta}$, $\operatorname{dist}(x, F_m) \le \lambda_m^\delta$ so that $\operatorname{dist}(x, F_m) \le C2^{-\alpha\delta m}$ for infinite values of m. If $\delta > \beta\gamma/s - \epsilon$, we get (6.3). When $s \ge \beta\gamma$, we observe that if $\eta_m > 0$ is an arbitrary sequence such that $\liminf \eta_m = 0$ and if

$$x \in \bigcap_{m \ge 1} F_m + [-\eta_m, \eta_m],$$

then $x \in K_\alpha$. This is because K_α is a compact set, and if $x \notin K_\alpha$, then $\operatorname{dist}(x, K_\alpha) = \eta > 0$ so that $\operatorname{dist}(x, F_m) \ge \eta$; hence we have a contradiction. Condition (6.3) is thus equivalent to $x \in K_\alpha$ as soon as $s \ge \beta\gamma$. The proof of the second part of the lemma is similar.

LEMMA 6.5. *The Hausdorff dimension of $E_{\alpha,\beta}$ is $1/\alpha\beta$. If $\gamma \le s \le \beta\gamma$, the Hausdorff dimension of $E^{(s)}$ is $s/\alpha\beta\gamma$. If $s > \beta\gamma$, the set $E^{(s)}$ is empty.*

The set $E_{\alpha,\beta}$ is defined by

$$E_{\alpha,\beta} = \bigcap_{m \ge 1} E_{\alpha,\beta}^{(m)} \quad \text{where } E_{\alpha,\beta}^{(m)} = G_m^{(\beta)} \cup G_{m+1}^{(\beta)} \cup \cdots \text{ and } G_m^{(\beta)} = F_m + [-\lambda_m^\beta, \lambda_m^\beta].$$

For any $\epsilon > 0$, we can cover $E_{\alpha,\beta}$ by the intervals I_q that appear in $G_n^{(\beta)}$, $n \ge m$. For a fixed n, there are 2^n such intervals of length $\sim 2^{-\alpha n\beta}$ so that if $d > 1/\alpha\beta$, $\sum |I_q|^d \le C$, where C does not depend on ϵ. Thus the Hausdorff dimension of $E_{\alpha,\beta}$ is bounded by $1/\alpha\beta$.

Now suppose that $\gamma \le s \le \beta\gamma$. Then

$$E^{(s)} = \left(\bigcap_{\delta < \frac{\beta\gamma}{s}} E_{\alpha,\delta} \right) \Bigg\backslash \left(\bigcup_{\delta > \frac{\beta\gamma}{s}} E_{\alpha,\delta} \right) \qquad \text{if } \gamma \le s < \beta\gamma,$$

while if $s = \beta\gamma$,

$$E^{(s)} = K_\alpha \Bigg\backslash \left(\bigcup_{\delta > 1} E_{\alpha,\delta} \right).$$

Checking is done the same way in both cases, so we suppose that $\gamma \le s < \beta\gamma$. Thus $E^{(s)} \subset E_{\alpha,\delta}$ for all $\delta < \beta\gamma/s$ so that $\dim(E^{(s)}) \le s/\alpha\beta\gamma$. Hence we have the two upper bounds for the Hausdorff dimensions in Lemma 6.5.

In order to obtain the lower bounds, we use a standard procedure. We construct a probability measure μ that is supported on $E_{\alpha,\beta}$ and has certain "scalings."

We now construct this measure.

Let $m_1 < m_2 < \cdots$ be an increasing sequence of integers that tends to ∞ quickly enough that for any $n \ge 1$, $m_{n+1} \ge \exp(m_n)$, and now let

$$K_{(\alpha,\beta)} = \bigcap_{n \ge 1} \tilde{G}_{m_n}^{(\beta)} \tag{6.8}$$

with

$$\tilde{G}_m^{(\beta)} = F_m + [-2^{-[\alpha\beta m]}, 2^{-[\alpha\beta m]}]. \tag{6.9}$$

This means that $\tilde{G}_m^{(\beta)}$ is a finite union of dyadic intervals, and the dyadic intervals that form $\tilde{G}_{m+1}^{(\beta)}$ will either be disjoint of those composing $\tilde{G}_m^{(\beta)}$ or included in them (just because they are dyadic intervals of smaller length). We have $[\alpha\beta m] \geq \beta[\alpha m]$, and the set $\tilde{G}_m^{(\beta)}$ is included in $G_m^{(\beta)}$ so that $K_{(\alpha,\beta)} \in E_{\alpha,\beta}$.

Let N_n be the number of intervals of length $2.2^{-[\alpha\beta m_n]}$ that can be found in $H_n = \tilde{G}_{m_1}^{(\beta)} \cap \cdots \cap \tilde{G}_{m_n}^{(\beta)}$, and let μ_n be the probability measure which on each of these N_n intervals takes the value $2^{[\alpha\beta m_n]}(2N_n)^{-1}dx$. We can easily check that $\mu_n \rightharpoonup \mu$ when $n \to \infty$, where $\mu = \mu_{(\alpha,\beta)}$ is supported by $K_{(\alpha,\beta)}$.

LEMMA 6.6. *There exists C such that $\forall I$ of length $|I| \leq 1/2$,*

$$\mu(I) \leq C|I|^{1/\alpha\beta} \log \frac{1}{|I|}. \tag{6.10}$$

Proof. We first estimate N_n. H_n is composed of N_n intervals of length $2^{-[\alpha\beta m_n]}$. When constructing H_{n+1}, we split each of these intervals into $2^{m_{n+1}-\beta m_n+\epsilon_n}$ intervals, where $|\epsilon_n| \leq 2$. Thus $N_{n+1} = N_n 2^{m_{n+1}-\beta m_n+\epsilon_n}$.

Now let I be an interval and define n by $2^{-\alpha\beta m_n} \leq |I| < 2^{-\alpha\beta m_{n-1}}$.

Consider the two cases $2^{-\alpha\beta m_n} \leq |I| < 2^{-\alpha m_n}$ and $2^{-\alpha m_n} \leq |I| < 2^{-\alpha\beta m_{n-1}}$. In the first one, I intersects at most two of the intervals that compose H_n so that

$$\mu(I) \leq CN_n^{-1} \leq C' 2^{-m_n+O(m_{n-1})} \leq C|I|^{1/\alpha\beta} \log(|I|)$$

since $m_n \geq \exp(m_{n-1})$. In the second case, suppose that $|I| \sim 2^{-\alpha j}$. Thus $\beta m_{n-1} \leq j \leq m_n$. I meets at most 2^{m_n-j} intervals so that $\mu(I) \leq 2^{m_n-j}/N_n$, but $N_n = N_{n-1}2^{m_n-\beta m_{n-1}+\epsilon_n}$. Thus

$$\begin{aligned}\mu(I) &\leq \frac{2^{-j}2^{\beta m_{n-1}-\epsilon_n}}{N_{n-1}} \leq C|I|^{1/\alpha\beta}2^{j/\beta}2^{-j}2^{\beta m_{n-1}}2^{-m_{n-1}+O(m_{n-2})}\\ &\leq C|I|^{1/\alpha\beta}2^{\frac{(\beta-1)}{\beta}(\beta m_{n-1}-j)+O(m_{n-2})}\end{aligned}$$

so that $\mu(I) \leq C|I|^{1/\alpha\beta}\log(|I|)$. Hence we have Lemma 6.6. □

We now prove the lower bounds in Lemma 6.5. We use the following slight modification of Hausdorff measure. Let $A \subset \mathbb{R}^m$ and R_ε be the set of all coverings of A by sets of diameter at most ε. Let

$$M(\varepsilon, d) \;=\; \inf_{r\in R_\varepsilon} \sum_{A_i \in r} (\operatorname{diam} A_i)^d \log\left(\frac{1}{(\operatorname{diam} A_i)}\right)$$

and let

$$d - \operatorname{mes}(A) \;=\; \limsup_{\varepsilon\to 0}\; M(\varepsilon, d)$$

be this "modified" d-dimensional Hausdorff measure. Of course, this modification does not change the Hausdorff dimension of A, which is

$$D \;=\; \inf\{d : d - \operatorname{mes}(A) = 0\} =\; \sup\,\{d : d - \operatorname{mes}(A) = +\infty\}.$$

We conclude with the following classical proposition (cf. [12]).

PROPOSITION 6.7. *Let $\mathcal{H}^s$ be the modified Hausdorff measure of dimension s. Let μ be a probability measure on $\mathbb{R}^m$, $F \in \mathbb{R}^m$. If $\limsup_{r\to 0} \mu(B(x,r))/r^s \log(1/r) < C \;\forall x \in F$,*

$$\mathcal{H}^s(F) \geq \frac{\mu(F)}{C}.$$

The first lower bound in Lemma 6.5 is thus a consequence of Lemma 6.6 and Proposition 6.7. The $\mathcal{H}^{1/\alpha\beta}$ measure of $E_{\alpha,\beta}$ is strictly positive, and thus the Hausdorff dimension of $E_{\alpha,\beta}$ is at least $1/\alpha\beta$.

We show that $\dim(E^{(s)}) \geq s/\alpha\beta\gamma$. Let μ be the probability measure $\mu_{\alpha,\beta\gamma/s}$. We check that

$$E^{(s)} \supset E_{\alpha,\beta\gamma/s} \Bigg\backslash \left(\bigcup_{\delta > \beta\gamma/s} E_{\alpha,\delta} \right) \tag{6.11}$$

and

$$\mu(E_{\alpha,\delta}) = 0 \quad \text{for any } \delta > \frac{\beta\gamma}{s};$$

since the union of these sets can be written as a countable union, the measure of their union vanishes so that the measure of $E^{(s)}$ is the same as the measure of $E_{\alpha,\beta\gamma/s}$, which is strictly positive. Hence we have the last point of Lemma 6.5.

We now prove the general case in Proposition 6.1.

Let $E_1, E_2, \ldots$ be disjoint subsets of $\mathbb{R}$ and suppose that $E_k \subset [a_k, b_k]$, where the $[a_k, b_k]$'s are disjoint. Let d_k be the Hausdorff dimension of E_k. Then the Hausdorff dimension of $\bigcup_{k\geq 1} E_k$ is $\sup(d_k)$.

We return to the function $F_{(\alpha,\beta,\gamma)}$. Clearly, $F_{(\alpha,\beta,\gamma)}$ has fast decay and is C^∞ outside of a compact set. After replacing $F_{(\alpha,\beta,\gamma)}(x)$ by $F_{(\alpha,\beta,\gamma)}(px+q)$, we can, without changing the spectrum of $F_{(\alpha,\beta,\gamma)}$, suppose that it is C^∞ outside any given interval $[a,b]$. Let $F_k(x) = f_{(\alpha_k,\beta_k,\gamma_k)}(x)$ be a sequence of functions as in Proposition 6.2 and consider the corresponding spectra $d_k(s)$. We can suppose that the singular supports of the $F_k(x)$'s are included in $[2^{-k-1}, 2^{-k}]$. (The singular support of a function is the closure of the set where this function is not C^∞.) We can also replace F_k by $\epsilon_k F_k$, where $\epsilon_k > 0$ tends to 0. Then let $G_1 = \sum_0^\infty \epsilon_k F_k$, and $d(s)$ is the supremum of the $d_k(s)$'s. The function G_1 thus constructed satisfies the requirements of Proposition 6.1, since we can easily check that a supremum of a countable set of functions of the form $ax1_{[b,c]}(x)$ is also a supremum of functions of the form $a1_{[b,c]}(x)$.

The construction of G_2 is now very easy. We remark that at each level j, the number of nonvanishing wavelet coefficients of $F_{(\alpha,\beta,\gamma)}$ is $o(2^j)$. Thus the same property holds for G_1 itself if we have chosen the contraction factors p (defined above) to be large enough. We now consider a function G_2 that has at each level j the same nonvanishing wavelet coefficients as G_1 but situated at different dyadic intervals. We group them in the smallest possible interval I_j centered at the origin. Thus the quantity (4.2) is the same for G_1 and G_2 so that these two functions share the same $B_p^{s,\infty}$ norm and hence the same function η. Nonetheless, if $x \neq 0$, there are a finite number of nonvanishing wavelet coefficients in a certain interval centered at x because the length of I_j tends to 0. Thus F_2 is C^∞ at x.

We now check that G_1 and G_2 are counterexamples to the following problem raised in [9]: Is η or ζ the Legendre transform of $m - d(\alpha)$?

Consider the function F defined by (6.1). At each level $j = [\alpha\beta m]$, it has 2^m wavelet coefficients equal to $2^{-(\gamma+1/2)j}$ so that for this j,

$$\left(\sum_k |C_{j,k}|^p \right)^{1/p} \sim 2^{(-(\gamma+1/2)+1/p\alpha\beta)j}$$

so that

$$\|F\|_{B_p^{s,\infty}} \sim 2^{js}2^{j(\frac{1}{2}-\frac{1}{p})}\left(\sum_k |C_{j,k}|^p\right)^{1/p} \sim 2^{j(s-\gamma+\frac{c\gamma}{p}-\frac{1}{p})}$$

and $\eta(p) = ap + 1 - ca$.

Thus $\eta(p)$ is linear and does not depend on b so that it clearly can be the Legendre transform of neither $cs1_{[a,b]}(s)$ (when $a \neq b$) nor the function 0. Thus in general, neither F_1 nor F_2 satisfies that η or ζ is the Legendre transform of its spectrum.

7. Acknowledgments. The author is thankful to Alain Arneodo, Emmanuel Bacry, Benoit Mandelbrot, and Yves Meyer for many enlightening discussions.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1978.

[2] A. Arneodo, E. Bacry, J. F. Muzy, *The thermodynamics of fractals revisited with wavelets*, Phys. A, 213 (1995), pp. 232–275.

[3] E. Bacry, A. Arneodo, J. F. Muzy, *Singularity spectrum of fractal signals from wavelet analysis: Exact results*, J. Statist. Phys., 70 (1993), p. 314.

[4] R. Benzi, L. Biferale, and G. Parisi, *On intermittency in a cascade model for turbulence*, Phys. D, 65 (1993), pp. 163–171.

[5] G. Brown, G. Michon, and J.Peyrière, *On the multifractal analysis of measures*, J. Statist. Phys., 66 (1992), pp. 775–790.

[6] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier–Stokes equations*, Comm. Pure Appl. Math., 35 (1982), p. 771.

[7] P. Collet, J. Lebowitz, and A. Porzio, *The dimension spectrum of some dynamical systems* J. Statist. Phys., 47 (1987), pp. 609–644.

[8] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[9] I. Daubechies and J. Lagarias, *On the thermodynamic formalism for multifractal functions*, in The State of Matter, M. Aizenman and H. Araki, eds., World Scientific, Singapore, 1994, pp. 213–265.

[10] A. Deliu and B. Jawerth, *Geometrical dimension versus smoothness*, Constr. Approx., 2 (1992), pp. 211–222.

[11] G. Eyink, *Besov spaces and the multifractal hypothesis*, J. Stat. Phys., 78 (1995), pp. 353–375.

[12] I. Falconer, *Fractal Geometry,* John Wiley, New York, 1990.

[13] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood–Paley theory and the study of function spaces*, CBMS Ser. 79, AMS, Providence, RI, 1991.

[14] U. Frisch and G. Parisi, *Fully developed turbulence and intermittency*, in Proc. Enrico Fermi International Summer School in Physics, North–Holland, Amsterdam, 1985, pp. 84–88.

[15] M. Holschneider, *Fractal wavelet dimensions and localization*, Comm. Math. Phys., 160 (1994), pp. 457–473.

[16] J. Hutchinson, *Fractals and self-similarity*, Indiana Univ. Math. J., 30 (1981), pp. 713–747.

[17] S. Jaffard, *Pointwise smoothness, two-microlocalization and wavelet coefficients*, Pub. Mat., 35 (1991), pp. 155–168.

[18] S. Jaffard, *Sur la dimension de Hausdorff des points singuliers d'une fonction*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1991), pp. 31–36 (in French).

[19] S. Jaffard, *Construction de fonctions multifractales ayant un spectre de singularités prescrit*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 19–24 (in French).

[20] S. Jaffard, *Formalisme multifractal pour les fonctions* C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 745–750 (in French).

[21] S. Jaffard and Y. Meyer, *Wavelet methods for Pointwise regularity and local oscillations of functions*, Mem. Amer. Math. Soc., 123 (1996).

[22] L. Olsen, *A multifractal formalism*, Adv. Math., to appear.

[23] B. Mandelbrot, *Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier*, J. Fluid Mech., 62 (1974), p. 331.

[24] C. Meneveau and K. R. Sreenivasan, *Measurement of $f(\alpha)$ from scaling of histograms and applications to dynamical systems and fully developed turbulence*, Phys. Lett. A, 137 (1989), pp. 103–112.

[25] Y. Meyer, *Ondelettes et opérateurs*, Hermann, Paris, 1990.

[26] J. F. Muzy, A. Arneodo, and E. Bacry, *A multifractal formalism revisited with wavelets*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), p. 245.

[27] S. M. Nicolskii, *Extension of functions of several variables preserving differential properties*, Amer. Math. Soc. Trans. (2), 83 (1969), pp. 159–188.

SIAM J. Math. Anal.
Vol. 28, No. 4, pp. 971–998, July 1997

MULTIFRACTAL FORMALISM FOR FUNCTIONS PART II: SELF-SIMILAR FUNCTIONS*

S. JAFFARD†

Abstract. In this paper we introduce and study the *self-similar functions.* We prove that these functions have a concave spectrum (increasing and then decreasing) and that the different formulas that were proposed for the multifractal formalism allow us to determine either the whole increasing part of their spectrum or a part of it. One of these methods (the wavelet-maxima method) yields their complete spectrum.

Key words. multifractal formalism, self-similarity, wavelet transform

AMS subject classifications. Primary, 26A15; Secondary, 76F99

PII. S0036141095283005

1. Introduction.

We proved in Part I that the multifractal formalism always yields an upper bound for the Hölder spectrum. Hence a very natural and important question arises: When does it yield the exact spectrum? Partial results exist for multifractal measures (especially when multinomial or for invariant measures of dynamical systems; see [6] and [8]). Hence we have similar results for the functions that are indefinite integrals of these measures. Apart from these functions, no general results hold. A few examples have been worked out: the scaling functions ϕ that appear in wavelet constructions [12], Riemann's nondifferentiable function [20], and the "peano-type" function of Polya [21]. These three examples exhibit a common feature: their graphs follow locally some self-similarity conditions. This is explicitly stated below in (2.7) for Riemann's function. The recursive definition of Polya's function is an exact self-similarity condition, and this is also the case for the scaling equation of the ϕ functions.

We can thus infer that the multifractal formalism probably holds if the function considered exhibits some kind of self-similarity. Of course, this assertion is very vague, and we are now far from guessing what is the weakest form of local self-similarity that implies the validity of the multifractal formalism. Our purpose is to verify it for a case study, i.e., under some restrictive assumptions for the self-similarity conditions. These assumptions are listed in Definition 2.1 below. In our opinion, this partial result is interesting for two reasons: 1) it is the first proof of the validity of the multifractal formalism for a several-parameter family of functions different from indefinite integrals of measures; 2) we also expect the methods we introduce to extend to more general settings. There is already some evidence for this. Since the first preprint version of this paper, Daubechies showed that some of the results concerning scaling functions can be deduced from our study [11]. Slimane showed that our restricive conditions concerning the contractions S_i in Definition 2.1 can be weakened [4].

2. Basic properties of self-similar functions.

In this section, we recall the definition of self-similar functions, give some examples, and derive their basic properties.

* Received by the editors March 13, 1995; accepted for publication (in revised form) April 5, 1996. http://www.siam.org/journals/sima/28-4/28300.html

† CMLA, École Normale Supérieure Cachan, 61 Avenue du Président Wilson, 94235 Cachan cedex, France and Département de Mathématiques, Université Paris XII, Faculté des Sciences et Technologie, 61 Avenue du Gal. de Gaulle, 94010 Créteil cedex, France (jaffard@cmla.ens-cachan.fr).

In the following two sections, we obtain the exact regularity of these functions at any point when the functions considered have uniform minimal regularity.

In section 4, we deduce these functions' Hölder spectra in the aforementioned case.

In section 5, we prove the validity of the multifractal formalism.

In section 6, we study the wavelet-maxima method. We show that after a slight modification, this method yields the complete spectrum, including the part where the infimum in the Legendre transform is obtained for negative values of q.

In section 7, we consider the more general case of unbounded self-similar functions.

We first recall the definition of self-similarity that was established in Part I.

Definition 2.1. *A function $F : \mathbb{R}^m \to \mathbb{R}$ is self-similar (of order $k \in \mathbb{R}^+$) if the following three conditions hold:*

- *There exists a bounded open set Ω and $S_1, \dots, S_d$ contractive similitudes such that*

$$S_i(\Omega) \subset \Omega, \tag{2.1}$$

$$S_i(\Omega) \cap S_j(\Omega) = \emptyset \quad \text{if } i \neq j. \tag{2.2}$$

(The S_i's are the product of an isometry with the mapping $x \to \mu_i x$, where $\mu_i < 1$.)

- *There exists a C^k function g such that g and its derivatives of order less than k have fast decay and F satisfies*

$$F(x) = \sum_{i=1}^{d} \lambda_i F(S_i^{-1}(x)) + g(x), \tag{2.3}$$

where the λ_i's are real or complex numbers.

- *The function F is not uniformly C^k in a certain closed subset of Ω.*

The first condition was first introduced by Hutchinson (in [17]) in order to study self-similar sets; it is called the "open-set condition." A stronger condition is sometimes required, namely,

$$S_i(\bar{\Omega}) \cap S_j(\bar{\Omega}) = \emptyset \quad \text{if } i \neq j; \tag{2.4}$$

this is called the "separated open-set condition."

Concerning the last point of the definition, if k is an integer, the condition must be understood as follows. Once restricted to a closed subset A of Ω, the derivatives of order $k-1$ of F do not belong to the Zygmund class. Thus for any $k \in \mathbb{R}^+$, this condition is equivalent to the existence of sequences $a_n \to 0$, $b_n \in A$, and $C_n \to \infty$ such that

$$|C(a_n, b_n)| \geq C_n a_n^k. \tag{2.5}$$

(This condition is a straightforward consequence of the wavelet characterization of the spaces $C^s(\mathbb{R}^m)$ that we recall below and of the localization of the wavelets.)

We will see that solutions of (2.3) need not necessarily be functions but can be distributions.

Recall the following notations introduced in Part I. Let

$$\tilde{Z}(a, q) = \int_{\mathbb{R}^m} |C(a, b)|^q db.$$

Then

$$\eta(q) = \liminf \frac{\log \tilde{Z}(a,q)}{\log a}.$$

Let

$$\alpha_{\min} = \inf_{i=1,\dots,d} \left(\frac{\log \lambda_i}{\log \mu_i} \right), \qquad \alpha_{\max} = \sup_{i=1,\dots,d} \left(\frac{\log \lambda_i}{\log \mu_i} \right).$$

We use this notation because $\alpha_{\min}$ will turn out to be the smallest pointwise Hölder regularity exponent of F and $\alpha_{\max}$ will be the largest (lower than k). Let τ be the function defined by

$$\sum_{i=1}^{d} \lambda_i^a \mu_i^{-\tau(a)} = 1.$$

The results concerning the multifractal formalism for self-similar functions are summed up in the following theorems.

THEOREM 2.2. *Suppose that F is self-similar. If $\alpha_{\min} > 0$, the function $d(\alpha)$ vanishes outside the interval $[\alpha_{\min}, \alpha_{\max}] \cup [k, +\infty)$ and is analytic and concave on $[\alpha_{\min}, \alpha_{\max}]$. Its maximal value $d_{\max}$ satisfies*

$$\sum \mu_i^{d_{\max}} = 1.$$

Let α_0 be the value for which this maximum is attained. First, suppose that g is C^∞. If $\alpha \le \alpha_0$, $d(\alpha)$ can be obtained by computing the Legendre transform of $\eta(q) - m$.

If g is only C^k, let p_0 be defined by $\eta(p_0) = kp_0$ and let $\alpha_1 < \alpha_0$ be the value of the inverse Legendre transform of $\eta(q) - m$ at p_0. If $\alpha \le \alpha_1$, $d(\alpha)$ can be obtained by computing the Legendre transform of $\eta(q) - m$.

Without any assumption on $\alpha_{\min}$, if $\sum |\lambda_j| \mu_j^m < 1$, the same results hold if we replace $d(\alpha)$ with $D'(\alpha)$, the packing dimension of the wavelet α-singularities (or by $D(\alpha)$, if g and the λ_i's are positive, and furthermore if the separated open-set condition holds).

The corresponding results concerning the *wavelet-maxima method* will be stated and proved in section 6 (see Theorem 2.3).

Before beginning our study of self-similar functions, we consider a few examples.

(1) *Indefinite integrals of multinomial measures in dimension* 1.

Let μ be a probability measure supported by $[0,1]$ and suppose that for any interval I,

$$\forall i = 1, \dots, d, \quad \mu(S_i(I)) = \lambda_i \mu(I)$$

with $\sum \lambda = D1i = 1$, the S_i's as above, and $\Omega = (0,1)$. Let

$$F(x) = \left(\int_0^x d\mu \right) - x \quad \forall x \in [0,1].$$

F vanishes at 0 and 1, and is smooth outside the intervals $S_i([0,1])$. One immediately checks that F is continuous and

$$\forall x \in S_i([0,1]), \quad F(x) = \lambda_i F(S_i^{-1}(x)) + g_i(x)$$

with g_i linear. Thus F is self-similar.

For any probability measure μ on $\mathbb{R}$, the scaling index of μ at x_0 is the supremum of all values of α such that

$$\exists C > 0, \forall \varepsilon > 0, \quad \mu([x_0 - \varepsilon, x_0 + \varepsilon]) \leq C\varepsilon^{\alpha}.$$

We can easily check that μ has a scaling index α at x_0 if and only if its indefinite integral F defined by $F(x) = \mu([0, x])$ is C^{α} at x_0 (see [2] or [19]). This property allowed Arneodo, Bacry, and Muzy [2] to determine the Hölder spectrum of the indefinite integrals of multinomial measures when the separated open-set condition holds. This remark shows that when F is the indefinite integral of a one-dimensional measure, some results derived in this paper are a consequence of similar results concerning measures (for $\alpha \in [0, 1]$) obtained by Brown, Michon, and Peyrière in [6]. Thus we are particularly interested in the case of functions that are not in bounded variation (BV), in which case Theorem 2.2 cannot be derived from corresponding results for measures.

(2) *Some self-similar fractal sets.* Consider, for instance, the example of the Van Koch set. Since it is a curve, it can be parametrized (in infinite ways) as the image of a mapping $t \to (x(t), y(t))$ from $[0, 1]$ to $\mathbb{R}^2$. This curve has dimension $\log 4/\log 3$ and a corresponding finite nonzero Hausdorff measure. Therefore, a *canonical* parametrization maps intervals of same length on sets of equal Hausdorff measure. The reader will immediately check that with *this* parametrization the Van Koch function is self-similar. Another example is supplied by Polya's function, a continuous mapping defined on $[0, 1]$ whose graph fills the area of a triangle. However, the lack of regularity of the function g in this case requires a specific treatment, and we plan to study the local regularity of this function in a forthcoming paper.

(3) *Lacunary trigonometric series and Riesz products.* Let

$$F_{\alpha}(x) = \sum_{j=0}^{\infty} 2^{-\alpha j} \sin 2\pi 2^j x$$

for $x \in [0, 1]$ and $0 < \alpha \leq 1$. Define

$$\begin{aligned} g(x) &= \sin 2\pi x \quad \text{if } \ x \in [0,1] \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Obviously,

$$F_{\alpha}(x) = 2^{-\alpha} F_{\alpha}(2x) + 2^{-\alpha} F_{\alpha}(2x - 1) + g(x)$$

so that the F_{α}'s are self-similar.

Another example is very similar. Consider the Riesz products

$$F_{\alpha,k}(x) = \prod_{j=1}^{\infty} (1 + 2^{-\alpha j} \sin(k^j x)),$$

where $0 < \alpha < 1$ and $k \in \mathbb{N}$, $k \geq 2$. It is a simple exercise to check that $\log(F_{\alpha,k}(x))$ is self-similar (the function g being $C^{\inf(2\alpha,1)}$). Since $F_{\alpha,k}(x)$ is bounded from above and below by strictly positive constants, $F_{\alpha,k}(x)$ and its logarithm share the same function η and the same spectrum so that the results that will be proved for self-similar functions will also hold for the Riesz products $F_{\alpha,k}(x)$.

(4) *Several dimension examples.* In dimension 1, the "geometry" contained in the transforms S_i is poor. In several dimensions, this is sometimes no longer the case. Consider two examples. First, if

$$\Omega = [-1,1]^2, \quad \text{let } i,j \in \left\{\frac{1}{2}, -\frac{1}{2}\right\}, \quad \text{and} \quad S_{i,j}(x) = \frac{1}{2}x + (i,j).$$

The $S_{i,j}$'s map the square Ω on its four subsquares of half size. If the homothety had a ratio smaller than 1/2, by iterating the $S_{i,j}$'s, we would get a kind of two-dimensional Cantor set. There exist more "exotic" examples. For instance, if

$$S_1^{-1} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad S_2^{-1} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} + (1,0),$$

the S_i's map a "fractal dragon" on their two self-similar components (see [9] or [16]).

In order to understand the scope and limitations of the model given by self-similar functions, it is interesting to mention a few classical examples of functions that, though not self-similar in the sense that we gave, satisfy functional equations that have similarities with (2.3).

First, the scaling function of the function φ used in the construction of orthonormal wavelet bases satisfies (see [10] or [12])

$$\varphi(x) = \sum a_k \varphi(2x-k),$$

but condition (2.2) does not hold except for some nonsmooth functions φ, such as characteristic functions of sets, in which case there exist examples similar to the fractal dragon that we mentioned above (see [9] or [16]). In [12], Daubechies and Lagarias recently proved that a converse formula to the multifractal formalism holds for these functions (i.e., the function $\eta(p)$ is the Legendre transform of $d(\alpha)$).

Our second example is the Brownian bridge on $[0,1]$. It satisfies

$$B(t) = \frac{1}{\sqrt{2}}B_1(2t) + \frac{1}{\sqrt{2}}B_2(2t-1) + \xi\Lambda(x), \tag{2.6}$$

where B_1 and B_2 are two Brownian bridges that have the same law as B, ξ is a Gaussian, $\Lambda(x) = \sup(x, 1-x)$ on $[0,1]$, and the three terms of the right-hand side of (2.6) are independent. We are in a situation where (2.3) holds "in law." Actually, the self-similar processes studied in, for instance, [3] also verify (2.3) "in law." We will not recall the definition of these processes here but only mention that in dimension 1, they coincide with the fractional Brownian motions. The reader can easily check that the results that we give below easily extend to this probabilistic setting. However, such results would be poor for the following reason. Direct methods yield sharper results for the the pointwise regularity of these processes and from a "multifractal point of view," the spectrum of these processes is not interesting since it vanishes everywhere except at one point.

Our last example is *Riemann's nondifferentiable function*

$$\Phi(x) = \sum_1^\infty \frac{1}{n^2}\sin(\pi n^2 x),$$

which was shown by Duistermaat in [13] to satisfy some functional equations similar to (2.3). For instance

$$\Phi(1+x) = \frac{\pi i}{12} - \frac{x}{2} + e^{i\pi/4}x^{3/2}\Big(4\Phi\Big(\frac{-1}{4x}\Big) - \Phi\Big(\frac{-1}{x}\Big)\Big) + \xi(x), \tag{2.7}$$

where ξ is a smooth function. Also, using the periodicity of Φ, a whole collection of similar equations can be derived. We are in a situation close to (2.3) but where the S_i's are not linear. (See [20] for an extension of the multifractal formalism to this case.)

We now determine the sense in which (2.3) has solutions, and we examine some basic regularity properties of these solutions. They will depend upon the assumptions that we make on the λ_i.

Iterating (2.3), we obtain for any N that

$$\begin{aligned} F(x) = & \sum_{n=0}^{N-1} \sum_{(i_1,\ldots,i_n)} \lambda_{i_1} \cdots \lambda_{i_n} g\big(S_{i_n}^{-1} \ldots S_{i_1}^{-1}(x)\big) \\ & + \sum_{(i_1,\ldots,i_N)} \lambda_{i_1} \cdots \lambda_{i_n} F\big(S_{i_N}^{-1} \cdots S_{i_1}^{-1}(x)\big)\,, \end{aligned} \tag{2.8}$$

so a (formal) solution of (2.3) is given by

$$F(x) = \sum_{n=0}^{\infty} \sum_{(i_1,\ldots,i_n)} \lambda_{i_1} \cdots \lambda_{i_n} g\big(S_{i_n}^{-1} \cdots S_{i_1}^{-1}(x)\big)\,. \tag{2.9}$$

Here F is written as a superposition of similar structures at different scales, reminiscent of some possible models of turbulence [5], [15], [26]. This formula also looks like a wavelet decomposition (except that g has no cancellation), and our proof of Proposition 3.2 in the next section will be similar to classical proofs of the regularity of wavelets series; see [18] or [22].

When the (formal) series (2.9) converges in a certain function space, a solution of (2.3) exists in this space. (Actually, it is easy to check that (2.9) converges almost everywhere if the separated open-set condition holds.) We will be particularly interested in three cases: first, when (2.3) has solutions that are locally L^1 functions; second, when the solutions have some global C^α smoothness (this case is important because it is the setting where the multifractal formalism works without any modification); and third, in spaces of distributions where the series converge when we make no assumption on the λ_i's. A good setting to study this last case is supplied by the real Hardy spaces, whose definition we now recall.

Suppose that we use an orthonormal basis of wavelets indexed by dyadic cubes. We denote these wavelets by ψ_λ and the corresponding wavelet coefficients by C_λ. The *real Hardy $\mathcal{H}^p$ space* (cf. [24]) is the set of distributions whose wavelet coefficients satisfy

$$\int \Big(\sum_{j,k} |C_\lambda|^2\, 2^{nj}\, \chi_\lambda(x) \Big)^{p/2} dx < +\infty, \tag{2.10}$$

assuming that the wavelets that we use are $C^{m(p^{-1}-1)}$ and have vanishing moments up to order $m(p^{-1}-1)$. This is a direct generalization of the space L^p when $p < 1$.

Recall that

$$\alpha_{\min} = \inf_{j=1,\ldots,d} \frac{\log|\lambda_j|}{\log \mu_j} \quad \text{and} \quad \alpha_{\max} = \sup_{j=1,\ldots,d} \frac{\log|\lambda_j|}{\log \mu_j}. \tag{2.11}$$

Proposition 2.3. *Suppose that* $\sum |\lambda_j|\mu_j^m < 1$. *In this case,* (2.3) *has a unique distribution solution, which is an* L^1 *function and is given by the series* (2.9). *Furthermore, if* $0 < \alpha_{\min} < k$, *this function is* $C^{\alpha_{\min}}$.

Suppose that $\sum |\lambda_j|\mu_j^m \geq 1$; *in that case,* (2.3) *may have several distribution solutions. Let* $p < 1$ *such that* $\sum |\lambda_j|^p\mu_j^m < 1$. *If* g *is* C^k *with* $k > m(p^{-1} - 1)$ *and if the moments of* g *of order less than* k *vanish,* (2.9) *converges in the Hardy real space* $\mathcal{H}^p$ *so that* (2.3) *has at least one solution in this space of distributions.*

Furthermore, these results are optimal.

Before proving Proposition 2.3, we begin with some preliminary results concerning the geometry of the mappings S_i. If A is a subset of $\mathbb{R}^n$, we define the mapping S by

$$S(A) = \bigcup_{i=1}^{d} S_i(A)$$

and let K be the set defined by

$$K = \bigcap_{n \in \mathbb{N}} S^n(\bar{\Omega}).$$

K is called the invariant compact set of S. Its Hausdorff dimension is $d_{\max}$ (defined in Theorem 2.2). We introduce some notation. Let i be a finite sequence $i = (i_1, \dots, i_n)$. We define $x_i = S_{i_1} \cdots S_{i_n}(0)$, and if the sequence i is infinite, $x_i = \lim_{n\to\infty} x_{(i_1,\dots,i_n)}$. Similarly, let $\mu_i = \mu_{i_1} \cdots \mu_{i_n}$ and $\lambda_i = \lambda_{i_1} \cdots \lambda_{i_n}$. Thus with each sequence $i \in \{1, \dots, d\}^{\mathbb{N}}$ we associate a unique point x_i in K. This correspondence is, in general, not one to one. (Consider, for instance, the example of lacunary trigonometric series where the dyadic points are the limit of two sequences.) However, the correspondence is clearly one to one if the separated open-set condition holds.

The points of K can also be represented as the limit points of the branches of the following tree T constructed in the "time-scale half-space." The treetop is conventionally the point $(0, 1) \in \mathbb{R}^m \times \mathbb{R}^+$. This treetop is linked to the d first nodes, which are the $(S_j(0), \mu_j)$'s. This point $(S_j(0), \mu_j)$ is linked to $(S_jS_k(0),\ \mu_j\mu_k), \dots$. If $\mathbb{R}^m$ is identified to $\mathbb{R}^m \times \{0\}$, then clearly the branch indexed by a sequence $i \in \{1, \dots, d\}^{\mathbb{N}}$ approaches the point x_i (and it is the only one which does so if the mapping $i \to x_i$ is one to one). This tree is related to the wavelet transform of F more precisely in Proposition 4.1. We will show that the order of magnitude of the wavelet transform of F near (x_i, μ_i) is $|\lambda_i|$.

Definition 2.4. *Let* $x \in \mathbb{R}^m$. *A "D-branch over* x*" is a branch of the tree of length* n *that starts from the origin* $(0, 1)$, *ends at*

$$(S_{i_1} \cdots S_{i_n}(0), \mu_{i_1} \cdots \mu_{i_n}),$$

and is such that

$$|S_{i_1} \cdots S_{i_n}(0) - x| \leq D\mu_{i_1} \cdots \mu_{i_n}.$$

When $D \leq 10\mathrm{Diam}(\Omega)$, *such a branch is a "main branch over* x*."*

This requirement means that the endpoint of the branch is—in the time-scale half-space—in a certain cone of width D over x. We often identify a branch with the sequence i that indexes it.

We will need the following lemma, which estimates the number of D-branches over a point x.

Lemma 2.5. *Let $x \in K$ and let $B_{j,D}(x)$ be the set of D-branches $(i_1, \ldots, i_n)$ over x such that*

$$2^{-j} \le \mu_{i_1} \cdots \mu_{i_n} < 2.2^{-j}. \tag{2.12}$$

The cardinality of this set of branches is bounded independentely of x and j by CD^m.

Proof. We can assume that the $S_{i_1} \cdots S_{i_n}(\Omega)$'s are disjoint. If not, the open-set condition implies that the corresponding sequences $(i_1, \ldots, i_n)$ and $(i'_1, \ldots, i'_m)$ satisfy—if $n \le m$, for instance—$(i_1, \ldots, i_n) = (i'_1, \ldots, i'_n)$. (One of the branches is included in the other.) In that case, we keep the longest sequence, dividing the cardinality of $B_{j,D}(x)$ by at most an absolute constant (which depends only on the values of $\mu_1, \ldots, \mu_d$). Thus we can assume that the $S_{i_1} \cdots S_{i_n}(\Omega)$'s are disjoint and are all included in $B(x, CD2^{-j})$, so if $B_{j,D}(x)^\#$ denotes the cardinality of $B_{j,D}(x)$,

$$B_{j,D}(x)^\# 2^{-mj} \mathrm{vol}(\Omega) \le C(D.2^{-j})^m,$$

and thus $B_{j,D}(x)^\#$ is bounded by CD^m. Hence Lemma 2.5 follows. □

We now prove Proposition 2.3. Existence and uniqueness in the L^1 case are straightforward. The last term in (2.8) tends to zero in L^1, so that (2.9) is the only (possible) solution in L^1, and it is actually in L^1 because the L^1 norm of series (2.9) is bounded by

$$C \sum_{|i| \le n} |\lambda_i| \mu_i^m = C \sum_{l \le n} \left(\sum_{j=1}^{d} |\lambda_j| \mu_j^m \right)^l \le C.$$

We estimate the C^s norm of F using the Littlewood–Paley characterization of this norm. For the reader's convenience, we recall this characterization.

Let ψ be a function in the Schwartz class whose Fourier transform vanishes outside $1 \le |\xi| \le 8$ and is equal to 1 on $2 \le |\xi| \le 4$. Let $\psi_l(x) = 2^{ml}\psi(2^l x)$. A function F belongs to C^s if and only if $|F * \psi_l(x)| \le C 2^{-sl}$.

We return to Proposition 2.3. We first split F as a sum $F = \sum F_j$, where F_j is series (2.9) restricted to the indices $i \in I_j$ such that

$$2^{-j} \le \mu_i < 2.2^{-j}.$$

Let $\omega_{l,j} = F_j * \psi_l$. If $l \ge j$, because of the localization and cancellation of ψ, for any N,

$$|\omega_{l,j}(x)| \le C_N \sum_{i \in I_j} \frac{\lambda_i 2^{-k(l-j)}}{(1 + 2^j |x - x_i|)^N}.$$

Because of Lemma 2.5, as soon as $N > m$,

$$\sum_{i \in I_j} \frac{1}{(1 + 2^j |x - x_i|)^N} \le C$$

so that

$$|\omega_{l,j}(x)| \le C \sup_{i \in I_j} |\lambda_i| 2^{-k(l-j)}.$$

If $j > l$, $|\omega_{l,j}(x)| \leq C \sup |F_j(x)|$ so that $|\omega_{l,j}(x)| \leq C \sup_{i \in I_j} |\lambda_i|$. Since $\sup_{i \in I_j} |\lambda_i| \leq C 2^{-\alpha_{\min} j}$, summing up, we obtain that $|(F * \psi_l)(x)| \leq C 2^{-\alpha_{\min} l}$. Hence we have the Hölder regularity of F.

In order to show that F belongs to $\mathcal{H}^p$, first notice that the regularity and cancellation that we requested for g is consistent with the atomic definition of $\mathcal{H}^p$ so that series (2.9) can be interpreted as a "vaguelette" decomposition of F (see [24]). Thus—following [24]—the "$\mathcal{H}^p$ norm" of F_j is bounded by

$$C \left(\int \left(\sum_{i \in I_j} |\lambda_i|^2 1_{|x - x_i| \leq \mu_i}(x) \right)^{p/2} dx \right)^{1/p} = C \left(\sum_{i \in I_j} \lambda_i^p \mu_i^m \right)^{1/p}.$$

By the same argument as in the L^1 case, this quantity is exponentially decreasing with j so that F belongs to $\mathcal{H}^p$.

The optimality of Proposition 2.3 can easily be checked via some explicit examples. The optimality of the global Hölder regularity is shown by example (2) above concerning lacunary trigonometric series. We sketch how to obtain the optimality of the L^1 and $\mathcal{H}^p$ criteria.

Let g be supported on $[1, 2]$ and suppose that F satisfies

$$F(x) = \lambda F(2x) + g(x).$$

If $\int g(x)dx \neq 0$ and $\lambda \geq 2$, series (2.9) does not converge in L^1 (or in any distribution space). If g has vanishing moments and $\lambda \geq 2$, the "$\mathcal{H}^p$ norm" of F can be calculated. For instance, when g is the function ψ that generates an orthonormal basis of compactly supported wavelets, ψ is properly contracted in order to be supported on the interval $[1, 2]$.

PROPOSITION 2.6. *If x does not belong to K, F is C^k in a neighborhood of x.*

Proof. Let α be such that $|\alpha| \leq k$, and let us show that the series

$$\sum \lambda_i \partial^\alpha (g \circ S_i^{-1}(.))$$

converges uniformly in a neighborhood of x.

This series is bounded in modulus by

$$\sum \frac{|\lambda_i| \mu_i^{-|\alpha|}}{(1 + \mu_i^{-1}|x - x_i|)^N} \leq \sum_j \sum_{i \in I_j} \frac{C 2^{C'j}}{(1 + 2^j |x - x_i|)^N},$$

but since $x \notin K$, $|x - x_i| \geq C > 0$ so that

$$\sum_{i \in I_j} \frac{C}{(1 + 2^j |x - x_i|)^N} \leq C' 2^{-(N-m)j}.$$

Choosing N large enough, we obtain Proposition 2.6 when $k \in \mathbb{N}$. The verification when k is not an integer is just as easy and is thus left to the reader. □

We conclude this section with a study on the uniqueness of solutions of (2.3). First, note that (2.8) holds for any N and that outside K the second term in (2.8) tends to 0 in C^k so that any distribution solution of (2.3) outside K is a function that satisfies (2.8). Thus if (2.3) has two solutions, their difference is a distribution supported by K, which is a solution of the homogeneous equation

$$F = \sum_{j=1}^{d} \lambda_j F \circ S_j^{-1}. \tag{2.13}$$

Since such a distribution is compactly supported, it belongs to a space $L^{p,s}$ (perhaps for a negative s). Note that $\|F \circ S_j^{-1}\|_{L^{p,s}} = \mu_j^{m/p-s} \|F\|_{L^{p,s}}$. Thus (2.13) implies that

$$\|F\|_{L^{p,s}} \leq \left(\sum_{j=1}^{d} |\lambda_j| \mu_j^{m/p-s} \right) \|F\|_{L^{p,s}},$$

and it has a nonvanishing solution in $L^{p,s}$ only if $\sum_{j=1}^{d} |\lambda_j| \mu_j^{(m/p)-s} > 1$.

Suppose that $\sum_{j=1}^{d} |\lambda_j| \mu_j^{m} < 1$. For all $s < 0$, let $p_0 m/(m-s)$. Then $\sum_{j=1}^{d} |\lambda_j| \mu_j^{(m/p)-s} < 1$ if $p < p_0$ so that (2.13) has no solution in $L^{p,s}$ for $p < p_0$ (hence for any p since F is compactly supported). Hence we have the uniqueness result in Proposition 2.3.

If $\sum_{j=1}^{d} |\lambda_j| \mu_j^{(m/p)-s} > 1$, it is easy to find distributions supported by K and solutions of (2.13). A trivial example is the Dirac mass at the origin, a solution of $\delta(.) = 2^m \delta(2.)$, but multinomial measures, such as the canonical measure on the triadic Cantor set, satisfy such equations. (The self-similar measures supported on K that we construct in section 4 also satisfy such equations.) In the case where $\sum_{j=1}^{d} |\lambda_j| \mu_j^{(m/p)-s} > 1$, we thus have no unique solution of (2.3), and we call (2.9) the *fundamental solution.*

For a given branch indexed by $i = (i_1, \ldots, i_n)$, let

$$\alpha(i) = \frac{\mathrm{Log}|\lambda_i|}{\mathrm{Log}\mu_i} \tag{2.14}$$

and denote the set $B_{j,10diam(\Omega)}(x)$ by $B_j(x)$.

In the next two sections, we prove the following result, which yields the exact regularity of f at each point of K when $\alpha_{\min} > 0$. (Recall that by definition f is Γ^α at x if α is the supremum of all β such that $f \in C^\beta(x)$.)

Proposition 2.7. *Suppose that* $\alpha_{\min} > 0$. *Let* $x \in K$. *Then* F *is* $\Gamma^{\alpha(x)}$ *at* x, *where*

$$\alpha(x) = \liminf_{j \to \infty} \inf_{i \in B_j(x)} \frac{\mathrm{Log}|\lambda_i|}{\mathrm{Log}\mu_i}. \tag{2.15}$$

The lower bound for $\alpha(x)$ will be obtained in section 3, and the upper bound will be obtained in section 4. In section 5, we determine the dimension of the set where F is Γ^α for a given α.

A case of special interest is when the separated open-set condition holds. In that case, there is only one branch over x and $i \to x(i)$ is onto so that if $i = (i_1(x), \ldots, i_n(x), \ldots)$ is the only sequence such that $x_0 = x(i)$, (2.14) and (2.15) become

$$\alpha(x_0) = \liminf_{n \to \infty} \frac{\mathrm{Log}|\lambda_{i_1(x_0)}| \cdots |\lambda_{i_n(x_0)}|}{\mathrm{Log}\mu_{i_1(x_0)} \cdots \mu_{i_n(x_0)}}.$$

3. A lower bound for regularity. We will need the following lemma, which yields an estimate for the products $\lambda_{i_1}\cdots\lambda_{i_n}$ on D-branches.

LEMMA 3.1. *Let $\Lambda_j(x)=\sup_{i\in B_j(x)}|\lambda_i|$ and $L_j(x)=\sum_{l=1}^{j}\Lambda_l(x)2^{-A(j-l)}$, where $A>\alpha_{\max}$. Then*

$$\liminf_{j\to\infty}\frac{\operatorname{Log}(L_j(x))}{-j\log 2}=\liminf_{j\to\infty}\frac{\operatorname{Log}(\Lambda_j(x))}{-j\log 2}\left(=\liminf_{j\to\infty}\inf_{i\in B_j(x)}\frac{\operatorname{Log}(\lambda_i)}{\operatorname{Log}(\mu_i)}\right)$$

and $\forall x\in\mathbb{R}^m$, if $\mu_i\sim 2^{-j}$,

$$|\lambda_i|\le CL_j(1+2^j|x-x_i|)^A.$$

In this lemma, we do not make any assumptions on the λ_i's. Let us prove the first assertion. $L_j\ge\Lambda_j$, and if $n(=n(j))$ is such that $n\le j$ and $\Lambda_n(x)2^{-A(j-n)}=\sup_{l\le j}\Lambda_l(x)2^{-A(j-l)}$, then $L_n\le n\Lambda_n(x)$. $A>\alpha_{\max}$ so that $n(j)\to\infty$ when $j\to\infty$. Hence we have the first assertion.

We now prove the second assertion. First, if i is a main branch, $|\lambda_i|\le L_j$. Now suppose that i is not a main branch. Let $i=(i_1,\dots,i_n)$ and let l be the largest integer such that the subbranch $(i_1,\dots,i_l)$ is a main branch over x. Clearly, $2^l|x-x_i|\sim 10\operatorname{diam}(\Omega)$ and $\lambda_i\le 2^{A(j-l)}\Lambda_l$ (because all of the λ_j's are $<2^{\alpha_{\max}}$), so

$$|\lambda_i|\le L_j\frac{\Lambda_l}{L_j}\le L_j2^{A(j-l)}\le L_j(C2^j|x-x_i|)^A.$$

Hence Lemma 3.1 follows.

PROPOSITION 3.2. *Let $x_0\in K$. The function F is $C^\beta(x_0)$ for any $\beta<\alpha(x_0)$.*

Proof. Let $x\in K$ and $P(x-y)$ be the Taylor expansion of order $[\beta]$ of (2.9) at x. We first check that this Taylor expansion yields a convergent series.

Let α be a multiindex such that $|\alpha|<\beta$. We have to check that the series

$$\sum\frac{|\lambda_i|\mu_i^{-|\alpha|}}{(1+\mu_i^{-1}|x-x_i|)^N}\tag{3.1}$$

is convergent. We split this sum into the sets

$$I_{j,l}=\{i\in I_j\text{ and }2^l<\mu_i^{-1}|x-x_i|\le 2^{l+1}\}.$$

Because of Lemma 2.5, each term has about 2^{lm} elements, and because of Lemma 3.1, on this set $I_{j,l}$,

$$\lambda_i\le CL_j(1+2^l)^A,$$

so series (3.1) is bounded by

$$C\sum_{j,l}L_j\mu_i^{-|\alpha|}(1+2^l)^{A-N}2^{lm}\le C\sum_{j,l}2^{(|\alpha|-\beta)j}2^{l(m-N+A)}$$

(since for j large enough, $L_j\le C2^{-\beta j}$), which is bounded because N can be chosen arbitrarily large.

Let $Tg_x(x-y)$ be the Taylor expansion of g of order $[\beta]$ at point x, i.e.,

$$Tg_x(x-y)=\sum_{|\gamma|\le[\beta]}\frac{\partial^\gamma g(x)}{\gamma!}(x-y)^\gamma.$$

Let J such that $2^{-J} \leq |x-y| \leq 2.2^{-J}$. Using formula (2.8) but stopping the iteration on each branch at the first level such that $\mu_i \leq 2^{-J}$, we obtain

$$\begin{aligned}(3.2)\qquad & F(y) - P(x-y) \\ &= \sum_{j\leq J}\sum_{i\in I_j} \lambda_i \left(g(S_i^{-1}(y)) - Tg_x(S_i^{-1}(x-y))\right) \\ &\quad + \sum_{j=J}\sum_{i\in I_J} \lambda_i F(S_i^{-1}(y)) - \sum_{j>J}\sum_{i\in I_j} \lambda_i Tg_x(S_i^{-1}(x-y)).\end{aligned}$$

The third sum is bounded in modulus by

$$\begin{aligned} C & \sum_{|\gamma|\leq\beta}\sum_{j\geq J} |\lambda_i|\mu_i^{-|\gamma|}|x-y|^{|\gamma|}(1+\mu_i^{-1}|x-x_i|)^{-N} \\ &\leq C\sum_{|\gamma|\leq\beta} |x-y|^{|\gamma|}\sum_{l}\sum_{j\geq J} L_j 2^{Al}2^{|\gamma|j}2^{l(m-N)} \\ &\leq C\sum_{|\gamma|\leq\beta} |x-y|^{|\gamma|}2^{(|\gamma|-\beta)J} \leq C|x-y|^{\beta}\end{aligned}$$

(where we have again split the sum into the sets $I_{j,l}$).

Because of the localization of F, the second sum is bounded by $C\sup_{i\in I_J}|\lambda_i| \leq C2^{-\beta J}$.

We now consider the first sum in (3.2). We consider two cases. Let $D = |x-y|^{-\epsilon}$ for an arbitrarily small ϵ. First, suppose that

$$|x-x_i| \leq D2^{-j}.$$

For each j, the sum has about D^m terms, and using the mean-value theorem, the sum of the corresponding terms is bounded by

$$\begin{aligned} D^m & \sum_{j\leq J}\sum_{i} L_j(1+D)^A|x-y|^{[\beta]+1}\mu_i^{-[\beta]-1} \\ &\leq C|x-y|^{[\beta]+1}\sum_{j\leq J} 2^{-\beta j}2^{([\beta]+1)j}D^{m+A} \leq C|x-y|^{\beta}D^{m+A}.\end{aligned}$$

Hence we have the bound that we claimed if we take ϵ small enough.

Now suppose that $|x-x_i| > D2^{-j}$; then $|\lambda_i| \leq CL_j(1+2^j|x-x_i|)^A$. Applying Lemma 2.5 with $D = 2^j|x-x_i|$, the remaining sum is bounded by

$$C\sum_{|\gamma|\leq\beta}\sum_{j\leq J}\frac{L_j(1+2^j|x-x_i|)^A(1+2^j|x-x_i|)^m}{(1+2^j|x-x_i|)^N}|x-y|^{\gamma}2^{\gamma j}$$

$$\leq \sum_{|\gamma|\leq\beta}\sum_{j\leq J} 2^{-\beta j}2^{\gamma j}|x-y|^{\gamma}(1+2^j|x-x_i|)^{-N+m+A},$$

and we obtain the bound in this case since $2^j|x-x_i| > |x-y|^{-\epsilon}$. Hence Proposition 3.2 follows.

4. An upper bound of the pointwise Hölder exponent. We will bound the regularity of F at each point in K by estimating the size of the wavelet transform in a neighborhood of such a point. The wavelet transform of F satisfies a functional equation similar to (2.3), which will enable us to obtain this estimate. Let $C(a,b)$ be the wavelet transform of F and $\omega(a,b)$ be the wavelet transform of g.

PROPOSITION 4.1. *There exists $A>0$ such that $\forall x\in K$, $J\in\mathbb{N}$. There exists $j\in[J-A,J]$, a branch $b=(j_1,\dots,j_n)$ in $B_j(x)$, $a\sim 2^{-j}$, and $t\in\Omega$ such that*

$$|t-x|\le Ca \quad \textit{and} \quad |C(a,t)|\ge C\Lambda_j(x).$$

Note that in this proposition, we do not have to make any assumptions on the uniform regularity of F, and we will actually use the proposition in cases where F is unbounded. Nonetheless, let us first show that if F has some minimal uniform regularity, Proposition 2.7 follows. To this end, we first recall a relation between the regularity of F and the size of the wavelet transform given by the following results (see Part I). Suppose that $s>0$. If $F\in C^s(x_0)$,

$$|C_{a,b}(F)|\le Ca^s\left(1+\frac{|b-x_0|}{a}\right)^s. \tag{4.1}$$

Thus Proposition 4.1 together with (4.1) shows that F is not smoother than $C^{\alpha(x)}$ at x. Thus using Proposition 3.2, we will have proved Proposition 2.7.

Proof of Proposition 4.1. We first prove Proposition 4.1 with $j\in[(1-\epsilon)J,J]$ (for an arbitrarily small ϵ). Let $C(a,b)$ be the wavelet transform of F and let $\omega(a,b)$ be the wavelet transform of g. Using (2.8) but stopping the iteration on each branch when $\mu_i\sim 2^{-J}$, we obtain

$$C(a,t)=\sum_{j=1}^{J}\sum_{i\in I_j}\lambda_i\omega\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)+\sum_{i\in I_J}\lambda_i\, C\left(\frac{a}{\mu_i},S_i^{-1}(t)\right). \tag{4.2}$$

Let $y\in\Omega$ be a fixed point that will be determined later. Let $x\in K$ and let $b=(j_1,\dots,j_n)$ be a branch over x. Let $t=S_{j_1}=CA\cdots S_{j_n}(y)$. Then

$$|x-t|\le C\mu_{j_1}\cdots\mu_{j_n}.$$

Hence we have the first condition of Proposition 4.1.

We want to show that on the set S_j, the main term in (4.2) corresponds to the branch b. This is intuitively clear because all terms in the first sum decay like a^k because of the smoothness of g, and since F is smooth outside Ω, all terms in the second sum decay also like a^k except precisely the one corresponding to the branch b. We make this argument more precise. We first prove the following bound for the first sum in (4.2):

$$\sum_{i\in I_j}\left|\lambda_i\omega\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)\right|\le Ca^k2^{kj}L_j(t). \tag{4.3}$$

First, note that because of the smoothness and decay of g,

$$\forall N\ge 0,\quad |\omega(a,b)|\le\frac{C_Na^k}{(1+|b|)^N}.$$

Thus

$$\begin{aligned}\sum_{i\in I_j}\left|\lambda_i\omega\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)\right| &\leq Ca^k\sum_{i\in I_j}\frac{|\lambda_i|}{\mu_i^k(1+|2^j(t-x_i)|)^N}\\ &\leq Ca^k2^{kj}L_j(t)\sum_{i\in I_j}\frac{1}{(1+|2^j(t-x_i)|)^{N-A}}\\ &\leq Ca^k2^{kj}L_j(t).\end{aligned}$$

Hence we have (4.3). Thus

$$\sum_{j\leq J}\sum_{i\in I_j}\left|\lambda_i\omega\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)\right|\leq Ca^k\sum_{j\leq J}2^{kj}L_j(t).$$

Since $\sup(\log\lambda_i/\log\mu_i)<k$, this series grows exponentially so that the first term in (4.2) is bounded by $Ca^k2^{kJ}L_J(t)$.

We now estimate the second term in (4.2) when $i\neq b$. Recall that A is the closed subset of Ω where by assumption F is not uniformly C^k. Let $A_\epsilon=A+B(0,\epsilon)$, where ϵ is a constant small enough that $A_\epsilon\subset\Omega$. Thus outside A_ϵ,

$$|C(a,b)|\leq C\frac{a^k}{(1+|b|)^N}$$

so that

$$C\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)\leq C\left(\frac{a}{\mu_i}\right)^k\frac{1}{(1+2^j|t-x_i|)^N}.$$

Thus we obtain, as above,

$$\sum_{i\in I_J,\, i\neq b}\lambda_iC\left(\frac{a}{\mu_i},S_i^{-1}(t)\right)\leq Ca^k2^{kJ}L_J(t).$$

Finally, from (4.2), we get

$$\left|C(a,t)-\lambda_jC\left(\frac{a}{\mu_j},S_j^{-1}(t)\right)\right|\leq Ca^k2^{kJ}L_J(t). \tag{4.4}$$

We now estimate the term corresponding to the sequence b. Recall that the last condition in Definition 2.1 is equivalent to the existence of sequences $a_n\to0$, $b_n\in A$, and $C_n\to+\infty$ such that $|C(a_n,b_n)|\geq C_na_n^k$ so that

$$\left|\lambda_bC\left(\frac{a}{\mu_b},S_b^{-1}(t)\right)\right|\geq|\lambda_b|C_n\left(\frac{a}{\mu_b}\right)^k.$$

Recall that $\Lambda_j=\sup_{i\in B_j(x)}|\lambda_i|$. Choosing a branch for which this supremum (taken on a finite number of terms) is attained, we get for this branch that

$$\left|\lambda_bC\left(\frac{a}{\mu_b},S_b^{-1}(t)\right)\right|\geq\Lambda_J(x)C_n2^{kJ}a^k$$

so that

$$|C(a,t)| = 2^{kJ}a^k[C_n\Lambda_J(t) + R],$$

where $|R| \leq CL_J(t)$. We choose n such that $C_n \geq 2C$, which determines a value of $a_n = a/\mu_b$. If

$$\Lambda_j(t) \geq \frac{1}{2}L_J(t),$$

the proposition is proved with $j = J$. Otherwise, since

$$L_J(t) = \Lambda_J + 2^{-A}\Lambda_{J-1} + 2^{-2A}\Lambda_{J-2} + \cdots,$$

one of the terms $2^{-lA}\Lambda_{J-l}$ must be large. More precisely, there exists l such that

$$2^{-lA}\Lambda_{J-l} \geq \frac{1}{10l^2}\Lambda_J. \tag{4.5}$$

(If several values of l satisfy (4.5), we choose the smallest.) We can choose the corresponding branch in Proposition 4.1, and since $l = o(J)$, this implies the irregularity of F at x. We found points in the "cone above x" where the wavelet transform is large. The statement of Proposition 4.1 is more precise because we will need precise estimates on the wavelet transform everywhere in order to estimate the integrals of the wavelet transform needed in the multifractal formalism. We have to check that we can choose $l \leq C$. We first prove that $l \leq \epsilon J$. We have $\Lambda_{J-l} \leq 2^{-\alpha_{\min}(J-l)}$ and $\Lambda_J \geq 2^{-\alpha_{\max}J}$ so that if (4.5) holds,

$$2^{-lA}2^{-\alpha_{\min}(J-l)} \geq \frac{1}{10l^2}2^{-\alpha_{\max}J},$$

which implies that

$$l \leq \left(\frac{\alpha_{\max} - \alpha_{\min}}{A - \alpha_{\min}}\right)J.$$

Choosing A large enough, we have $l \leq \epsilon J$ for ϵ arbitrarily small. For this branch b,

$$|C(a,t)| \geq \frac{1}{2}\lambda_b(x)C\left(\frac{a}{\mu_b}, S_b^{-1}(t)\right) \tag{4.6}$$

and $C(a/\mu_b, S_b^{-1}(t)) \sim 1$. Hence we have Proposition 4.1 when $j \in [(1-\epsilon)J, J]$. We now want to prove that the proposition holds for $j \in [J-A, J]$.

Suppose that t is a point inside Ω such that the $S_i(t)$'s do not approach the boundary of Ω. We know that

$$C(a,t) = \omega(a,t) + \sum_{j=1}^{d}\lambda_j C\left(\frac{a}{\mu_j}, S_j^{-1}(t)\right),$$

but $|\omega(a,t)| \leq C_1a^k$ and outside of a certain neighborhood of Ω, $|C(a,t')| \leq C_2a^k$. Let

$$i = (i_1, \ldots i_n) \quad \text{and} \quad i' = (i_1, \ldots i_n, i_{n+1}).$$

Thus if t and r are such that $|C(\mu_i,t)-r\lambda_i|\le e$, then

$$\begin{aligned}|C(\mu_{i'},S_{i'}(t))-r\lambda_{i'}|&\left|\omega(\mu_{i'},S_{i'}(t))+\sum_j^d\lambda_jC\left(\frac{\mu_{i'}}{\mu_j},S_j^{-1}(t)\right)-r\lambda_{i'}\right|\\ &\le C_1(\mu_{i'})^k+\sum_{j\ne i_{n+1}}\lambda_jC\left(\frac{\mu_{i'}}{\mu_j}\right)^k+\lambda_{i_{n+1}}|r\lambda_i-C(\mu_i,t)|\\ &\le\left(C_1+C\sum_{j\ne i_{n+1}}\frac{\lambda_j}{(\mu_j)^k}\right)\mu_{i'}^k+e\lambda_{i+1}\le C\mu_{i'}^k+e\lambda_{i+1}.\end{aligned}$$

We start with a branch i such that $r\sim 1$ and $e=0$, which is possible because of the first part of the proof. After one iteration, we obtain an error of $C\mu_i^k$; after two iterations, we get an error of $C\mu_i^k\lambda_{i+1}+C\mu_{i+1}^k,\ldots$; and after j iterations, the error is

$$C\mu_i^k\left(\frac{\lambda_j}{\lambda_i}+\mu_i\frac{\lambda_j}{\lambda_{i+1}}+\cdots\right)\sim C\mu_i^k\frac{\lambda_j}{\lambda_i}$$

so that $|C(\mu_j,t')-r\lambda_j|\le C'\mu_i^k(\lambda_j/\lambda_i)\le\epsilon\lambda_j$, where t' is on the subtree deduced from t. Thus $C(\mu_j,t')\sim r\lambda_j$. Hence Proposition 4.1 follows.

Note that Propositions 3.2 and 4.1 show that the wavelet transform of F is "large" near the tree T, and thus the ramifications of this tree of wavelet maxima reflect the "dynamics" of self-similarity as stated by Arneodo, Bacry, and Muzy in [2].

It is remarkable that these results do not depend on the function g. If g were replaced by another function, the new F would have the same regularity at every point. Only the global smoothness of g is important. It defines a value beyond which one can no longer calculate the regularity of F.

5. Determination of the Hölder spectrum. In this section, we prove that for $\alpha<k$, the Hölder spectrum of a self-similar function is the Legendre transform of the function τ defined by

$$\sum_{i=1}^d\lambda_i^a\mu_i^{-\tau(a)}=1.$$

PROPOSITION 5.1. *Let $\alpha<k$ and define $d(\alpha)$ as the Hausdorff dimension of the set of points x where F is $\Gamma^\alpha(x)$. Then $d(\alpha)$ is given on $[0,k)$ by*

$$d(\alpha)=\left(\inf_a a\alpha-\tau(a)\right). \tag{5.1}$$

We will need the following proposition (Proposition 4.9 in [25]) in the proof of Proposition 5.1.

PROPOSITION 5.2. *Let $\mathcal{H}^s$ be the Hausdorff measure of dimension s. Let μ be a probability measure on $\mathbb{R}^m$, $F\subset\mathbb{R}^m$, and C be such that $0<C<+\infty$. Then*

- *if* $\limsup_{r\to 0}\mu(B(x,r))/r^s<C\ \forall x\in F,\ \mathcal{H}^s(F)\ge\mu(F)/C$;
- *If* $\limsup_{r\to 0}\mu(B(x,r))/r^s>C\ \forall x\in F,\ \mathcal{H}^s(F)\le 2^s/C$.

Proof of Proposition 5.1. Let $a\in\mathbb{R}$, $b=-\tau(a)$, and $P_i=\lambda_i^a\mu_i^b$. Thus $\sum P_i=1$. We first consider on K a probability measure ν such that

$$\forall(i_1,\ldots,i_n),\quad \nu(S_{i_1}\cdots S_{i_n}(K))=P_{i_1}\cdots P_{i_n}. \tag{5.2}$$

The construction of such a measure by induction is straightforward (see, for instance, [17]). Let $x \in K$, $s > 0$ and $r > 0$ and consider the set $B_j(x)$, where $2^{-j} \leq r < 2.2^{-j}$. Then

$$\frac{\nu(B_r(x))}{r^s} = \sum_{i \in B_j(x)} \frac{\lambda_i^a \mu_i^b}{\mu_i^s} \sim \sup_{i \in B_j(x)} \left(\lambda_i \, \mu_i^{\frac{b-s}{a}} \right)^a$$

(because the number of branches over x in $B_j(x)$ is bounded by an absolute constant).

Suppose that

$$\frac{b-s}{a} < -\alpha(x).$$

Then $\limsup_{r \to 0} \nu(B_r(x))/r^s \to +\infty$ so that, using Proposition 5.2, $\mathcal{H}^s(\Gamma^\alpha) = 0$. Thus $d(\alpha) \leq b + a\alpha$ so that $d(\alpha) \leq -\tau(a) + a\alpha \ \forall a \in \mathbb{R}$.

In order to prove Proposition 5.1, we have to show that the infimum is reached. Using Proposition 5.2, it is sufficient to find a and b such that $\nu(\Gamma^\alpha) > 0$.

Suppose that a and b are solutions of the following system

$$(5.3) \qquad \left.\begin{aligned} \sum_{i=1}^d \lambda_i^a \mu_i^b = 1, \\ \frac{\sum P_i \mathrm{Log}\lambda_i}{\sum P_i \mathrm{Log}\mu_i} = \alpha, \end{aligned}\right\}$$

where $P_i = \lambda_i^a \mu_i^b$. (In Lemma 5.3, we will determine the values of α for which this system has a solution.)

If $(i_1, \dots, i_n)$ is a branch over x, let $(n_j)_{j=1,\dots,d}$ be the proportion of $j's$ in the sequence $i_1, \dots, i_n$ and let F be the subset of K composed of the points x such that

$$(5.4) \qquad n_j \to P_j$$

(meaning here that $\forall \varepsilon > 0$, $\exists n : \forall m, \geq n$, if $(i_1, \dots, i_m)$ is a branch over x, then $|n_j - P_j| \leq \varepsilon \ \forall j = 1, \dots, d$ for this branch).

If $x \in F$, then

$$\liminf_{j \to \infty} \inf_{i \in B_j(x)} \frac{\mathrm{Log}\lambda_i}{\mathrm{Log}\mu_i} = \lim \frac{\mathrm{Log}\lambda_i}{\mathrm{Log}\mu_i} = \frac{\sum P_j \mathrm{Log}\lambda_j}{\sum P_j \mathrm{Log}\mu_j} = \alpha$$

so that $F \subset \Gamma^\alpha$.

Let ν be the corresponding probability defined by (5.2). We can associate with ν another probability P defined on $\{1, \dots, d\}^{\mathbb{N}}$ as follows. If $i = (i_1, \dots, i_n)$ and I_i is the subset of $\{1, \dots, d\}^{\mathbb{N}}$ of all of the sequences starting with $(i_1, \dots, i_n)$, then

$$P(I_i) = P_{i_1} \dots P_{i_n}.$$

With probability P, the i_n's are a sequence of i.i.d. random variables. The strong law of large numbers implies that with probability 1, $n_j \to P_j$ for a sequence $i \in \{1, \dots, d\}^{\mathbb{N}}$. Clearly, ν is the image of the probability P by the application $x(i)$. We want to show that on K, ν-almost everywhere $n_j \to P_j$. It would be obvious if $x(i)$ were one to one.

First, note that if $(i_1, \ldots, i_n)$ is a branch over x, so is $(i_1, \ldots, i_{n-1})$. Now suppose that (5.4) fails. For n arbitrarily large, we can find a branch over x such that

$$|n_j - P_j| \geq \varepsilon. \tag{5.5}$$

Consider such a sequence of branches over x for $n \to \infty$. Since at a scale r there are at most N branches over x, (following Lemma 2.5) such branches for which (5.4) fails can be grouped into at most N sets of increasing branches. Among these, at least one, $\tilde{b}_x$, has infinite length.

We call a branch of infinite length i such that $x = x(i)$ a *principal branch over* x. Because of Lemma 2.5, for each x, there are at most N such branches. Clearly, $\tilde{b}_x$ is a principal branch over x.

Consider the event $\{x$ is such that (5.4) fails$\}$. It is included in the event $\{\exists b$ principal branch over x such that (5.5) holds$\}$. Since the probability for one given branch is 0, the probability that (5.5) holds for at least one of the (at most) N principal branches over x is also 0 such that for probability ν, almost every point of K is such that (5.4) holds. Thus $\nu(F) = 1$, and since $F \subset \Gamma^\alpha, \nu(\Gamma^\alpha)1$. We can now apply Proposition 5.2. Hence we have the first part of Proposition 5.1.

LEMMA 5.3. *Suppose that* $\alpha_{\min} < \alpha_{\max}$. *System* (5.3) *has a solution if and only if*

$$\alpha_{\min} < \alpha < \alpha_{\max}. \tag{5.6}$$

If $\alpha_{\min} = \alpha_{\max}$, *the only solution is* $\alpha = \alpha_{\min} = \alpha_{\max}$.

Proof. One can easily check that if $a_1, \ldots, a_d,\ b_1, \ldots, b_d > 0$, and the P_i's are weights (i.e., $0 < P_i$ and $\sum P_i = 1$), then

$$\inf\left(\frac{a_i}{b_i}\right) \neq \sup\left(\frac{a_i}{b_i}\right) \implies \inf\left(\frac{a_i}{b_i}\right) < \frac{\sum P_i a_i}{\sum P_i b_i} < \sup\left(\frac{a_i}{b_i}\right)$$

so that (5.6) is necessary. Now suppose that this holds. Since $\sum \lambda_i^a \mu_i^{-\tau(a)} = 1$, $\forall i, \tau(a) \leq a\text{Log}\lambda_i/\text{Log}\mu_i$. If $a \to +\infty$, $\tau(a) \leq a\ \alpha_{\min}$ so that if j is such that $\text{Log}\lambda_j/\text{Log}\mu_j > \alpha_{\min}$, then $\lambda_j^a \mu_j^{-\tau(a)} \to 0$. Thus if $\alpha_{\min}$ is reached for i in a subset $J \subset \{1, \ldots, d\}$, then $\sum_{i \in J} \lambda_i^a \mu_i^{-\tau(a)} \to 1$, but $\sum_{i \in J} \lambda_i^a \mu_i^{-\tau(a)} = \sum_{i \in J} e^{\text{Log}\mu_i(a\alpha_{\min} - \tau(a))}$ so that $\tau(a)/a \to \alpha_{\min}$. Thus all of the $P_i \to 0$ except for $i \in J$ so that $\sum P_i \text{Log}\lambda_i / \sum P_i \text{Log}\mu_i \to \alpha_{\min}$.

If $a \to -\infty$, $\tau(a) \leq a\alpha_{\max}$ and the same argument yields $\sum P_i\text{Log}\lambda_i / \sum P_i \text{Log}\mu_i \to \alpha_{\max}$. By continuity, $\sum P_i \text{Log}\lambda_i / \sum P_i \text{Log}\mu_i$ takes all values between $\alpha_{\min}$ and $\alpha_{\max}$.

Notice that if $\alpha_{\min} = \alpha_{\max} = \alpha_0$, then $\alpha = \alpha_0$ is the only possible value for which (5.3) has a solution.

6. Proof of the multifractal formalism. Now that we have determined the spectrum of a self-similar function, we will prove the multifractal formalism for these functions. First, we will do so for the wavelet-transform integral method. We recall the formulas that are used. We compute

$$\tilde{Z}(a, q) = \int_{\mathbb{R}^m} |C(a, b)|^q db.$$

Let

$$\eta(q) = \liminf \frac{\log \tilde{Z}(a, q)}{\log a}. \tag{6.1}$$

The Hölder spectrum is computed using the formula

$$d(\alpha) = \inf_q (q\alpha - \eta(q) + m). \tag{6.2}$$

In order to estimate $\tilde{Z}(a,q)$ for self-similar functions, we first have to estimate $C(a,b)$ everywhere. Let

$$\begin{cases} i &= (i_1, \ldots, i_n), \\ \Omega_i &= S_{i_1} \cdots S_{i_n}(\Omega), \\ B_i &= \Omega_i + B(0,a), \\ C_i &= B_{(i_1,\ldots,i_{n-1})} - B_{(i_1,\ldots,i_n)}. \end{cases}$$

If $a \le \mu_i$, $\mathrm{Vol}(B_i) \sim (\mu_i)^m$ and $\mathrm{Vol}(C_i) \le C(\mu_i)^m$. Inequality (4.6) shows that there exists one point $b \in B_i$ and an a such that $C2^{-j} \le a \le 2^{-j}$ for which the order of magnitude of $C(a,b)$ is Λ_n. We show that this order of magnitude holds on a ball of size $\sim a$. To this end, we bound $C(a,b)$ in B_i (and also in C_i, which will be useful later).

LEMMA 6.1. *Let $a > 0$ and B_i be such that $a \sim \mu_i$. Then if $b \in B_i$,*

$$|C(a,b)| \le CL_i,$$

and if $a \le \mu_{i_1} \cdots \mu_{i_n}$, then if $b \in C_i$,

$$|C(a,b)| \le CL_i \left(\frac{a}{\mu_i}\right)^k.$$

Lemma 6.1 is derived from (2.8) exactly as in the beginning of the proof of Proposition 2.7. We leave the details to the reader.

We return to the estimation of $C(a,b)$. In order to prove that it keeps the same order of magnitude in a ball of size $\sim a$, we bound $\bigtriangledown_b C(a,b)$ and $\partial_a C(a,b)$. Let $\partial_b C(a,b)$ be a partial derivative of $C(a,b)$ in a certain direction $b_0 \in \mathbb{R}^m$. Clearly,

$$\partial_b C(a,b) = \frac{1}{a}\tilde{C}(a,b),$$

where $\tilde{C}(a,b)$ is a wavelet transform using the wavelet $\partial\psi$.

The bound given by Lemma 6.1 for $C(a,b)$ holds for $\tilde{C}(a,b)$ so that

$$|\partial_b C(a,b)| \le \frac{C}{a} L_i.$$

Since at a certain point of B_i, $C(a,b)$ is of the order of magnitude of L_i, this is also the case on a ball of size $\sim a$.

If we now differentiate the wavelet transform with respect to the variable a, the same procedure yields

$$\partial_a C(a,b) = \frac{1}{a}\tilde{C}(a,b),$$

where $\tilde{C}(a,b)$ is a wavelet transform using the wavelet $\psi(x) - x.\bigtriangledown\psi(x)$, so that $C(a,b)$ is of the order of magnitude of L_i on a ball of size $\sim a$ in the time-scale half-space.

Furthermore, on C_i,

$$|C(a,b)| \leq C\frac{a^k \lambda_i}{(\mu_i)^k}. \tag{6.3}$$

Let A_j be the interval $[2^{-(j+A)}, 2^{-j}]$. For each branch i such that $\mu_i \sim 2^{-j}$, Proposition 4.1 shows that there exists a ball of radius at least $C2^{-j}$ in the time-scale half-space located near x_i and in scale in the interval A_j, where $|C(a,b)| \geq C\lambda_i$. Thus

$$\begin{aligned} C\sum_{\mu_i \sim 2^{-j}} 2^{-j(m+1)}\lambda_i^q &\leq \int_{A_j\times\mathbb{R}^m} |C(a,b)|^q da\, db \\ &\leq C'\sum_{\mu_i\sim 2^{-j}} 2^{-j(m+1)}\lambda_i^q + O\left(2^{-j}\sum_{\mu_i\geq 2^{-j}} \frac{2^{-kqj}\lambda_i^q}{\mu_i^{qk-m}}\right) \end{aligned}$$

so that

$$\begin{aligned} C\sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^m\lambda_i^q &\leq 2^j\int_{A_j\times\mathbb{R}^m} |C(a,b)|^q da\, db \\ &\leq C'\left[\sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^m\lambda_i^q + O\left(\sum_{\mu_i\geq 2^{-j}} \frac{2^{-kqj}\lambda_i^q}{\mu_i^{qk-m}}\right)\right]. \end{aligned} \tag{6.4}$$

We first estimate the term

$$\sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^m\lambda_i^q. \tag{6.5}$$

The reader should notice that in the following estimation, we do not have to assume that q is positive. This remark will be useful in section 7.

Recall that $\tau(q)$ is such that

$$\sum_{j=1}^d \mu_j^{-\tau(q)}\lambda_j^q = 1. \tag{6.6}$$

Thus

$$\sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^m\lambda_i^q \sim 2^{-(m+\tau(q))j}\sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^{-\tau(q)}\lambda_i^q.$$

Let

$$F(j) = \sum_{2^{-j}\leq\mu_i<2.2^{-j}} \mu_i^{-\tau(q)}\lambda_i^q.$$

From (6.6), we get

$$\sum_{|i|=N} \mu_i^{-\tau(q)}\lambda_i^q = \left(\sum_{j=1}^d \mu_j^{-\tau(q)}\lambda_j^q\right)^N = 1$$

so that

$$\sum_{|i|\leq N_0} \mu_i^{-\tau(q)} \lambda_i^q = N_0. \tag{6.7}$$

Clearly,

$$\sum_{j=1}^{J} F(j) = \sum_{\mu_i \geq 2^{-J}} \mu_i^{-\tau(q)} \lambda_i^q. \tag{6.8}$$

After permuting the indexation of the S_i's, we can assume that $\mu_1 = \inf \mu_i$ and $\mu_d = \sup \mu_i$.

The right-hand side of (6.8) contains all of the terms of length N if $\mu_1^N \geq 2^{-J}$ and no terms of length M if $\mu_d^M \leq 2^{-J}$. Thus from (6.7) and (6.8), we get

$$N \leq \sum_{j=1}^{J} F(j) \leq M,$$

which can be written as

$$J \frac{\mathrm{Log} 2}{\mathrm{Log}(\frac{1}{\mu_1})} \leq \sum_{j=1}^{J} F(j) \leq J \frac{\mathrm{Log} 2}{\mathrm{Log}(\frac{1}{\mu_d})}$$

so that there exist $C_1, C_2 > 0$ such that

$$\left.\begin{aligned} \limsup \frac{F(j)}{j} &\leq C_1, \\ \limsup F(j) &\geq C_2. \end{aligned}\right\} \tag{6.9}$$

Thus

$$\limsup \frac{1}{j} 2^{(m+\tau(q))j} \sum_{2^{-j} \leq \mu_i < 2.2^{-j}} \mu_i^m \lambda_i^q \leq C$$

and

$$\limsup 2^{(m+\tau(q))j} \sum_{2^{-j} \leq \mu_i < 2.2^{-j}} \mu_i^m \lambda_i^q \geq C'.$$

Now consider the term

$$\sum_{\mu_i \geq 2^{-j}} \frac{2^{-kqj} \lambda_i^q}{\mu_i^{qk-m}}.$$

This is bounded by

$$C 2^{-kqj} \sum_{\mu_i \geq 2^{-j}} \lambda_i^q \mu_i^{m-qk}. \tag{6.10}$$

We split this sum into bands B_l defined by $2^{-l-1} \leq \mu_i < 2^{-l}$. Using (6.7), we get

$$C2^{-kqj} \sum_{\mu_i \in B_l} \lambda_i^q \mu_i^{m-qk} \leq C2^{-kqj} 2^{-l(m-kq+\tau(q))} \sum_{\mu_i \in B_l} \lambda_i^q \mu_i^{-\tau(q)}$$

$$\leq C2^{-kqj} 2^{-l(m-kq+\tau(q))}.$$

Now suppose that q is such that $m - kq + \tau(q) \leq 0$. Equation (6.10) is bounded by

$$C2^{-kqj} 2^{-(m-kq+\tau(q))j} \leq C2^{-(m+\tau(q))j},$$

and using (6.4) and (6.9), we obtain the followingproposition.

Proposition 6.2. *Suppose that F is self-similar and let q be such that*

$$\tau(q) \leq kq - m. \tag{6.11}$$

Then

$$\limsup_{a \to 0} a^{-m-\tau(q)} \int |C(a,b)|^q db \geq C > 0 \tag{6.12}$$

$$\limsup_{a \to 0} \frac{a^{-m-\tau(q)}}{|\log a|} \int |C(a,b)|^q db \leq C' < +\infty. \tag{6.13}$$

This result together with Proposition 2.7 proves the multifractal formalism for the wavelet-integral method.

The multifractal formalism is also valid for the structure function method since we showed in Part I that $\zeta(q) = \eta(q)$ for $q > 1$. However, the restriction $q > 1$ shows that it might not yield the whole left-hand side of the spectrum but a smaller part corresponding to the region where the infimum in the Legendre transform formula is attained for $q > 1$.

Note that if k can be chosen arbitrarily large (when g is C^∞), condition (6.11) reduces to $q \geq 0$. We consider the case of negative values of q in the next section.

7. The wavelet-box method.

In this section, we first show some pitfalls of the wavelet-maxima method and then show that a slight modification allows us to obtain the spectrum even for its decreasing part.

Let us first briefly recall the principles of the wavelet-maxima method. Consider for a given $a' > 0$ the local maxima of the function $b \to C(a', b)$. Generically, they belong to a line of maxima $b = l(a)$ defined in a small left-neighborhood $[a'', a']$ of a' by the following condition: $b \to C(a,b)$ has a local maximum for $b = l(a)$. The wavelet-transform maxima method requires first the computation of

$$Z(a,q) = \sum_l \sup_{b=l(a)} |C(a,b)|^q,$$

where l is a line of maxima of the wavelet transform defined on $[a'', a']$, and the sum is taken on all lines of local maxima defined in a left-neighborhood $[a'', a']$ of a'. We then define

$$\theta(q) = \liminf \frac{\log Z(a,q)}{\log a}.$$

The counterexamples concerning the wavelet-maxima method that were given in Part I easily adapt to the self-similar case. Suppose, for instance, that g is one of these counterexamples supported by the interval $[3,4]$. (We can make this assumption because they are compactly supported and these properties still hold after a contraction and a translation.) Then

$$F(x) = \sum 2^{-\alpha j} F(2^j x),$$

where $\alpha > 0$ is self-similar and yields a similar counterexample. Nonetheless, we will see that we can adapt the wavelet-maxima method so that it yields results as good as and even better than the other methods. To this end, we introduce a slight variant, the *wavelet-box method*.

Let C be a parameter larger than 1. The wavelet-box method consists of dividing $\mathbb{R}^m$ into cubes of length C and, for each cube included in Ω, keeping only the largest local maximum (if there is one on each of these cubes). Clearly, this procedure has the advantage of not taking into account accumulations of lines of local maxima, on which the counterexamples of Part I were based. We still use the notation $\theta(q)$, which will avoid confusion with the wavelet-maxima method.

THEOREM 7.1. *Under the same hypotheses as Theorem* 2.2*, the wavelet-box method yields the increasing part of the spectrum of self-similar functions. Furthermore, if* $d_{\max} = m$ *or, equivalently, if* $\overline{\cup S_i(\Omega)} = \overline{\Omega}$*, the wavelet-box method yields the whole spectrum of self-similar functions, provided that we keep only the largest maximum of a box of size* Ca *for a constant* C *large enough, i.e.,* $d(\alpha)$ *can be obtained by computing the Legendre transform of* $\theta(q) - m$*.*

With regard to the increasing part of the spectrum (corresponding to $p \geq 0$), the theorem is a consequence of the wavelet-integral method because of the following lemma.

LEMMA 7.2. *The two quantities*

$$\int |C(a,b)|^q db \quad \textit{and} \quad a^m \sum_{\max} |C(a,b)|^q$$

are of the same order of magnitude if the sum is computed as in the wavlet box method.

This result is quite straightforward since we estimated $\int |C(a,b)|^q db$ precisely by computing its order of magnitude near the wavelet maxima. We showed that its value is about λ_i^q near the tree T and smaller far from the tree. This shows that there exists at least one maximum near each point of the tree. The estimation of $a^m \sum_{\max} |C(a,b)|^q$ then follows exactly the estimation performed in Proposition 6.2. Thus, in that case, the verification of the fractal formalism reduces to the verification for the integral formula, and the multifractal formalism holds when using the two-wavelet methods. Note that for positive q's we do not have to restrict the sum to cubes included in Ω, which is interesting if we do not have a priori knowledge on Ω.

Before proving Theorem 2.3 for the decreasing part of the spectrum, we make some general remarks.

Consider again the case $q < 0$ but for the quantity $\sum_{\max} |C(a,b)|^q$. An important difference from the exact computations of [2] appears. Recall that the authors of [2] were interested in the case where F is the indefinite integral of a multinomial measure supported by a Cantor set. In this case, the wavelet maxima are situated on the "tree over the Cantor set" since F is piecewise constant outside this set, so, for a small enough, the wavelet transform vanishes there. Thus in this case, the last term of (6.4) vanishes, and the same proof as above shows that Proposition 6.2 will hold for $q < 0$ (with the same restriction concerning the distance between the maxima).

In the general case that we consider in this paper, F is a C^k function outside K for which we do not have special information (since g is C^k but arbitrary). Thus there may be extremely small wavelet maxima "far away" (on the scale of a) from K (i.e., at a distance $\gg a$). Thus no formula involving negative values of q can work reasonably. We present an example of this phenomenon.

First, note that if $\Omega \neq K$, it is easy to construct examples where g and thus F will be locally constant so that $F(x+h) - F(x)$ will vanish on an open subset for h small enough, as does $C(a,b)$ for a small enough. Thus $\zeta(q)$ and $\eta(q)$ take the value $+\infty$ for negative values of p so that in all generality, computing $\zeta(q)$ and $\eta(q)$ for negative values of p does not make sense. The same problem appears for formulas involving wavelet maxima. Of course, in the regions where $C(a,b)$ vanishes, there are no more maxima. However, it is easy to construct g with lines of very small maxima as follows.

Let ψ be a C^{k+2} function with moments of order $k+1$ vanishing and supported by $[3/2, 2]$, and let

$$h(x) = \sum_{j \geq 0} 2^{-2kj} \psi(2^{2j}x).$$

This is supported by $[0,2]$, and if ψ is the analyzing wavelet and $a = 2^{-j}$, $C(a,b)$ vanishes outside the interval $|b| \leq 2^{-j}/2$ (if $|b| \leq 2^{-j}/8$), but $C(a,0) = C2^{-kj}$. Thus $C(a,b)$ has a line of maxima that goes through the interval $|b| \leq 2^{-j}/2$ and the supremum on this line is of the order of magnitude of 2^{-kj}. Now suppose that $F = g$ in a neighborhood of 0 (which we can always assume if $\Omega \neq K$). Then $Z(a,q)$ is larger than Ca^{-kq}. We see that no bound of $Z(a,q)$ can be found independent of k. If the multifractal formalism held, since $d(\alpha)$ is independent of k, the order of magnitude of $Z(a,q)$ would be as well. Hence we have a contradiction.

We make one final remark on Theorem 2. First, suppose that g is C^∞. If $\alpha \leq \alpha_0$, the infimum in the Legendre transform is obtained for $q > 0$ (because $\tau(0) = -d_{\max}$ and τ is convex and increasing). In that case, we cannot directly calculate $d(\alpha)$ up to α_0 since we cannot use a C^∞ wavelet with all vanishing moments, but following [1], this can be done using a sequence of increasingly smoother wavelets and determining increasingly larger parts of the spectrum. The case where g is C^k is still easier to check.

We now want to show that using the wavelet-box method, we can recover the left part of the spectrum corresponding to negative values of q when

$$\overline{\cup S^j(\Omega)} = \Omega.$$

(This implies that there exists no region where F is smooth.) The validity of this condition can actually be checked on the part of the spectrum computed for $q > 0$ since at the maximum (the case where $q = 0$) $d(\alpha) = d_{\max}$, which satisfies

$$\sum \mu_j^{d_{\max}} = 1.$$

However, the condition $\overline{\cup S^j(\Omega)} = \Omega$ can be rewritten $\sum \mu_j^m = 1$ since $\mathrm{Vol}(\Omega_j) = \mu_j^m \mathrm{Vol}(\Omega)$. Thus $\overline{\cup S^j(\Omega)} = \Omega$ is equivalent to $d_{\max} = m$, which is easy to check.

In this case, the tree T leaves no void in the region of the upper half-plane above Ω. (By this we mean that for any (a,b) in this region, we can find a point of the tree in a domain $[a/C, Ca] \times [b - Ca, b + Ca]$); however, after perhaps increasing the constant C, we can also find a point $(\mu_i, S^i(t))$ where the order of magnitude of the wavelet transform of F is λ_i. Thus if we modify the wavelet-maxima method by imposing the condition that we first take the largest local maximum on the box $[a/C, Ca] \times [b - Ca, b + Ca]$ (which amounts to considering a kind of maximal function), we see that for a given scale a,

$$\sum_{\max} |C(a,b)|^q \sim \sum_{a \sim \mu_i} |\lambda_i|^q.$$

Returning to the estimation of (6.5), we see that

$$\limsup_{a \to 0} a^{-\tau(q)} \sum_{\max} |C(a,b)|^q db \geq C > 0,$$

$$\limsup_{a \to 0} \frac{a^{-\tau(q)}}{|\log a|} \sum_{\max} |C(a,b)|^q db \leq C' < +\infty.$$

Hence we have the multifractal formalism in that case.

Note that the constant C in the definition of the wavelet-box method must be chosen "large enough" depending on the self-similar function that is analyzed.

8. Unbounded self-similar functions. Thus far, we made the assumption $\alpha_{\min} > 0$ (which is equivalent to $|\lambda_j| < 1 \; \forall j = 1, \ldots, d$). This implied that $F \in C^{\alpha_{\min}}$. Thus we were interested only in Hölder exponents larger than $\alpha > 0$. However, we would like to consider negative exponents, which, as mentioned before, should be pertinent in some applications. We have already discussed the definition of negative exponents in Part I. We will now show that the multifractal formalism holds using a slightly different definition for the Hölder spectrum. We suppose that

$$\sum_{i=1}^{d} |\lambda_i| |\mu_i|^m < 1 \tag{8.1}$$

so that a function F that satisfies (2.3) belongs to L^1. Condition (8.1) can clearly give rise to unbounded functions. In fact, the following result holds.

LEMMA 8.1. *Suppose that one of the λ_j's satisfies $|\lambda_j| > 1$ and that g does not vanish identically. Then $\forall x \in K$, F is unbounded in any neighborhood of x so that $d(\alpha) = 0 \; \forall \alpha$.*

Proof. First, note that a straightforward estimation yields that if $F \in L^\infty$, then $|C(a,b)| \leq C$ so that if F is bounded in a neighborhood of x $|C(a,b)| \leq C$ if b is in a (perhaps) smaller neighborhood of x.

Let $x \in K$ and $i = (i_1, \ldots, i_n, \ldots)$ such that $x = x(i)$. Let j be such that $|\lambda_j| > 1$. Then we fix an n and define $i' = (i_1, \ldots, i_n, j, j, \ldots) = (i'_1, \ldots)$. Proposition 4.1 shows that there exists an x_0 such that if $b_l = S^{i'_1} \cdots S^{i'_l}(x_0)$ and $a_l = \mu_{i'_1} \cdots \mu_{i'_l}$, $|C(a_l, b_l)| \sim |\lambda_{i'_1} \ldots \lambda_{i'_l}|$ and is thus unbounded. If n is large enough and $l \geq n$, the points b_l are arbitrarily close to x so that F is unbounded in any neighborhood of x. Hence we have Lemma 8.1.

In this case, the Hölder spectrum of F for $\alpha > -n$ is trivial: $d(\alpha) = 0 \; \forall \alpha > -n$.

Note that if $\sup |\lambda_j| = 1$, F can be either unbounded or bounded depending on the specific values taken by the function g, as shown by the following example.

Suppose that F satisfies

$$F(x) = F(2x) + g(x)$$

and suppose that g is Lipschitz. If g is positive and does not vanish at 0, F is unbounded, whereas if $g(0) = 0$, F is bounded.

Lemma 8.1 suggests that if F is unbounded, the spectrum defined by the Hausdorff dimension of α-singularities is not the right quantity to consider, but we should instead compute the packing dimension of strong singularities.

We now suppose that the separated open-set condition holds and that g and the λ_j's are positive. We prove Theorem 2.2 in that case.

Let $x \in K$ and $i = (i_1, \dots, i_n, \dots)$ be the (unique) branch over x. We define two subsets A_1 and A_2 of $B(x, \mu_i)$ as follows. First, let Ω' be a set where $g(x) \geq C > 0$. Suppose that $\lambda_1 > 1$ and let $i' = (i_1, \dots, i_n, 1, \dots, 1)$, where the number p of ones will be made precise later. Since

$$F(y) = \sum \lambda_i g(S_i^{-1}(y)),$$

if $y \in S^{i'}(\Omega)$, $F(y) \geq C_1 \lambda_i \lambda_1^p$. $S^{i'}(\Omega)$ is the set A_1. If $\operatorname{dist}(y, K) \geq \mu_i/2$, $F(y) \leq C_2 \lambda_i$. A_2 is the set of points of $B(x, \mu_i)$ such that $\operatorname{dist}(y, K) \geq \mu_i/2$.

We choose p such that $C_1 \lambda_1^p \geq 2C_2$. The volumes of A_1 and A_2 are $\sim \mu_i^m$ (because of the separated open-set condition). Thus if

$$\beta(x) = \limsup \frac{\log |\lambda_{i_1}(x)| \cdots |\lambda_{i_N}(x)|}{\log \mu_{i_1}(x) \cdots \mu_{i_N}(x)}, \tag{8.2}$$

F has a strong singularity of order $\beta(x)$ at x. Furthermore, the order of magnitude of F on a subset of $B(x, C\mu_{i_1}(x) \cdots \mu_{i_N}(x))$ of size comparable to the size of this ball is exactly $\mu_{i_1}(x) \cdots \mu_{i_N}(x)$ so that if $\beta(x)$ is negative, the order of the strong singularity at x is not higher than $\beta(x)$. Hence we have the first part of the following proposition. (The last part is a direct consequence of Proposition 4.1.)

PROPOSITION 8.2. *If the separated open-set condition holds, if g and the λ_j's are positive, and if* (8.1) *holds, F has a "strong singularity" of order $\beta(x)$ at x but no strong singularity of higher order. Furthermore, without any assumption on g, the S_i's or the λ_i's, F has a "wavelet singularity" of order $\beta(x)$ at x.*

It would be interesting to prove this proposition without the assumption of the separated open-set condition. It clearly holds if $K \neq \Omega$ (i.e., if $\sum \mu_i^m \neq 1$) because in that case the choice that we made for the y's is still possible.

The case where $K = \Omega$ is perhaps less interesting for applications since f is then nowhere locally bounded (by Lemma 7.2), which is usually not realistic for "physical" functions. Actually, in the case of three-dimensional turbulence, the singularities seem to concentrate on a set of dimension < 3 (see [1] or [23]).

We now determine the packing dimension of the strong α-singularities of F.

PROPOSITION 8.3. *Under the same assumptions, the packing dimension of the strong α-singularities of f is given by $D(\alpha) = \inf(a\alpha - \tau(a))$.*

We already know that $D(\alpha) \leq \inf(a\alpha - \tau(a))$, so we have to prove only the lower bound. We will use the following result (see [14] or [25]).

PROPOSITION 8.4. *Let H^s be the packing measure of dimension s. Let μ be a probability measure on $\mathbb{R}^m$, $F \in \mathbb{R}^m$, and C be such that $0 < C < +\infty$.*

If $\liminf_{r \to 0} \mu(B_r(x))/r^s < C \ \forall\, x \in F$, $H^s(F) \geq \mu(F)/C$.

The proof of Proposition 8.3 follows the proof of Proposition 5.1 since we can take exactly the same measure μ and the same set F. It is now easy to check that the multifractal formalism holds in this setting because the proof of Proposition 6.2 actually holds without any change.

We conclude this paper with the proof of the following corollary, which shows that the multifractal formalism holds in a more general setting.

COROLLARY 8.5. *Let A be a pseudodifferential elliptic operator of order s whose symbol σ satisfies*

$$\forall\alpha,\ \beta,\quad |\partial_x^\alpha \partial_\xi^\beta \sigma(x,\xi)| \le C(\alpha,\beta)(1+|\xi|)^{s-\beta}.$$

Suppose that h is a self-similar function that satisfies

$$h(x) = \sum_{i=1}^{d} \lambda_i h(S_i^{-1}(x)) + g(x),$$

and if s is negative, suppose further that the moments of g of order at most $|s|$ vanish.

Let $F = A(h)$. If $F \in C^\epsilon(\mathbb{R}^m)$ for an $\epsilon > 0$, then its spectrum is a concave function whose increasing part is given by

$$d(\alpha) = \inf_q(\alpha q - \eta(q) + m).$$

Proof. Following a result of Calderón and Zygmund [7], A can be written $\tilde{A}(-\Delta)^{m/2}$ (up to a regularizing operator), where $\tilde{A}$ and $\tilde{A}^{-1}$ are Calderón–Zygmund operators. Clearly, $(-\Delta)^{m/2}h$ is a self-similar function if $m \ge 0$ and also if $m < 0$ and g has the corresponding number of vanishing moments. Thus the multifractal formalism holds for F because Calderón–Zygmund operators are continuous on the Besov spaces so that $(-\Delta)^{m/2}h$ and F share the same function η. Since these operators are also continuous on the two-microlocal spaces (see [18]), $(-\Delta)^{m/2}h$ and F have the same Hölder regularity at each point (perhaps up to a logarithmic correction).

REFERENCES

[1] A. Arneodo, E. Bacry, J. F. Muzy, *The thermodynamics of fractals revisited with wavelets*, Phys. A, 213 (1995), pp. 232–275.

[2] E. Bacry, A. Arneodo, J. F. Muzy, *Singularity spectrum of fractal signals from wavelet analysis: Exact results*, J. Statist. Phys., 70 (1993), p. 314.

[3] A. Benassi, S. Jaffard and D. Roux, *Elliptic Gaussian random processes*, Rev. Mat. Iberoamericana, to appear.

[4] M. Ben Slimane, Ph.D. thesis, Ecole Nationale des Parts et Chaussies, 1996.

[5] R. Benzi, L. Biferale, and G. Parisi, *On intermittency in a cascade model for turbulence*, Phys. D, 65 (1993), pp. 163–171.

[6] G. Brown, G. Michon, and J.Peyrière, *On the multifractal analysis of measures*, J. Statist. Phys., 66 (1992), pp. 775–790.

[7] A. P. Calderón and A. Zygmund, *Singular integral operators and differential equations*, Amer. J. Math., 79 (1957), pp. 901–921.

[8] P. Collet, J. Lebowitz, and A. Porzio, *The dimension spectrum of some dynamical systems*, J. Statist. Phys., 47 (1987), pp. 609–644.

[9] A. Cohen, Thèse Université Paris Dauphine, Paris, 1990.

[10] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[11] I. Daubechies, Personnal communication, 1993.

[12] I. Daubechies and J. Lagarias, *On the thermodynamic formalism for multifractal functions*, in The State of Matter, M. Aizenman and H. Araki, eds., World Scientific, Singapore, 1994, pp. 213–265.

[13] J. J. Duistermaat, *Selfsimilarity of Riemann's non-differentiable function,* Nieuw Arch. Wisk., 9 (1991), pp. 303–337.

[14] I. Falkoner, *Fractal Geometry,* John Wiley, New York, 1990.

[15] U. Frisch and G. Parisi, *Fully developed turbulence and intermittency,* in Proc. Enrico Fermi International Summer School in Physics, North–Holland, 1985, pp. 84–88.

[16] K. Grochenig and W. Madych, *Selfsimilar lattice tilings,* J. Fourier Anal. Appl., 1 (1994), pp. 131–171.

[17] J. Hutchinson, *Fractals and self-similarity,* Indiana Univ. Math. J., 30 (1981), pp. 713–747.

[18] S. Jaffard, *Pointwise smoothness, two-microlocalization and wavelet coefficients*, Publ. Mat., 35 (1991), pp. 155–168.

[19] S. Jaffard, *Formalisme multifractal pour les fonctions,* C.R. Acad. Sci. Paris Sér I Math., 317 (1993), pp. 745–750.

[20] S. Jaffard, *The spectrum of singularities of Riemann's function*, Rev. Mat. Iberoamericana, 12 (1996), pp. 441–460.

[21] S. Jaffard and B. B. Mandelbrot, *Local regularity of nonsmooth wavelet expansions and application to the Polya function*, Adv. Math., 120 (1996), pp. 265–282.

[22] S. Jaffard and Y. Meyer, *Wavelet methods for pointwise regularity and local oscillations of functions*, Mem. Amer. Math. Soc., 123 (1996).

[23] B. Mandelbrot, *Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier*, J. Fluid Mech., 62 (1974), p. 331.

[24] Y. Meyer, *Ondelettes et Opérateurs*, Hermann, Paris, 1990.

[25] C. Tricot, *Rectifiable and fractal sets*, in Fractal Geometry and Analysis, J. Bélair and S. Dubuc, eds., NATO ASI Ser., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991, pp. 367–404.

[26] V. Ziemin, *Hierarchic models of turbulence*, Izv. Atmos. Oceanic Phys., 17 (1981), pp. 941–949.

SIAM J. MATH. ANAL.
Vol. 28, No. 4, pp. 999–1014, July 1997

STABILITY AND ORTHONORMALITY OF MULTIVARIATE REFINABLE FUNCTIONS*

W. LAWTON†, S. L. LEE‡, AND ZUOWEI SHEN‡

Abstract. This paper characterizes the stability and orthonormality of the shifts of a multidimensional (M, c) refinable function ϕ in terms of the eigenvalues and eigenvectors of the transition operator $W_{c_{\text{au}}}$ defined by the autocorrelation c_{au} of its refinement mask c, where M is an arbitrary dilation matrix. Another consequence is that if the shifts of ϕ form a Riesz basis, then $W_{c_{\text{au}}}$ has a unique eigenvector of eigenvalue 1, and all of its other eigenvalues lie inside the unit circle. The general theory is applied to two-dimensional nonseparable (M, c) refinable functions whose masks are constructed from Daubechies' conjugate quadrature filters.

Key words. refinement equations, interpolatory refinable functions, dilation matrix, subdivision operators, transition operators

AMS subject classifications. 41A15, 41A30, 42C05, 42C15

PII. S003614109528815X

1. Introduction.

In this paper, we present a complete characterization of the stability and orthonormality of the shifts of a refinable function in terms of the refinement mask by analyzing the simplicity of eigenvalue 1 of the transition operator.

Denote by $\ell^1(\mathbf{Z^d})$ and $\ell^2(\mathbf{Z^d})$ the spaces of absolutely summable and modulus-square-summable complex-valued sequences defined on $\mathbf{Z^d}$, respectively. Let $M \in \mathbf{Z^{d\times d}}$ be a $d\times d$ integer matrix with eigenvalues of modulus > 1 and with $|\det M| = m > 1$. Let $c \in \ell^1(\mathbf{Z^d})$ and $\phi : \mathbf{R^d} \to \mathbf{C}$ be a complex-valued function. The equation

$$\phi(x) = \sum_{q\in\mathbf{Z^d}} mc(q)\phi(Mx - q) \tag{1.1}$$

is called a *refinement equation*. The matrix M is called a *dilation matrix*. The sequence c is called a *refinement mask*, and the function ϕ is called an (M, c) *refinable function* or (M, c) *scaling function*. We assume that $\int \phi(x)dx = 1$.

Denote by $L^1(\mathbf{R^d})$ and $L^2(\mathbf{R^d})$ the spaces of Lebesque-integrable and modulus-square-integrable functions defined on $\mathbf{R^d}$, respectively. The class of all tempered distributions on $\mathbf{R^d}$ will be denoted by $\mathcal{S}'$. The dilation operator M associated with the dilation matrix M is defined for all functions ϕ by $M\phi(x) := \phi(Mx),\ x \in \mathbf{R^d}$. This can be extended to all distributions $\phi \in \mathcal{S}'$ by defining

$$\langle M\phi, f\rangle := \frac{1}{m}\langle \phi, M^{-1}f\rangle \quad \text{for all } f \in \mathcal{S},$$

where $\mathcal{S}$ denotes the class of all infinitely differentiable functions with rapid decay at infinity. Similarly, the shift operator $T^p\phi(x) := \phi(x - p),\ p \in \mathbf{Z^d}$, for functions may be extended to distributions by

$$\langle T^p\phi, f\rangle := \langle \phi, T^{-p}f\rangle \quad \text{for all } f \in \mathcal{S}.$$

* Received by the editors June 22, 1995; accepted for publication (in revised form) March 12, 1996. This research was supported by the National University of Singapore Wavelets Programme.
http://www.siam.org/journals/sima/28-4/28815.html

† Institute of System Sciences, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511 (wlawton@iss.nus.sg).

‡ Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511 (matleesl@math.nus.sg, matzuows@leonis.nus.sg).

The refinement equation (1.1) may now be extended to include distributions $\phi \in \mathcal{S}'$ by writing

$$\phi = \sum_{q\in\mathbf{Z^d}} mc(q) M T^q \phi. \tag{1.2}$$

A distribution ϕ that satisfies (1.2) is called an (M, c) *refinable distribution.*

The Fourier transform of a sequence $a \in \ell^1(\mathbf{Z^d})$ will be denoted by $\widehat{a}$ and is defined by

$$\widehat{a}(u) := \sum_{p\in\mathbf{Z^2}} a(p) e^{-ipu},$$

where $i \equiv \sqrt{-1}$. Note that $\widehat{a}(u)$ is a complex-valued 2π-periodic continuous function on $\mathbf{R^d}$ and thus is defined on the d-dimensional torus $\mathbf{T^d}$. For a finitely supported sequence $(a_j)_{j\in\mathbf{Z^d}}$ with support in $[0, N-1]^d$, we define N as its *length.*

For any continuous function f defined on $\mathbf{R^d}$, we shall denote by $f_|$ the sequence $(f(p))_{p\in\mathbf{Z^d}}$, which is the restriction of f to $\mathbf{Z^d}$.

The Fourier transform of a function $f \in L^1(\mathbf{R^d})$ is

$$\widehat{f}(u) := \int_{\mathbf{R^d}} f(x) e^{-ixu} dx.$$

This maps $\mathcal{S}$ onto itself and extends to all tempered distributions $\mathcal{S}'$ by duality.

We shall assume throughout this paper that c is a finitely supported sequence that satisfies

$$\sum_{p\in\mathbf{Z^d}} c(p) = 1. \tag{1.3}$$

Then there exists a compactly supported (M, c) refinable distribution ϕ, unique up to a constant multiple, such that its Fourier transform admits the infinite-product representation

$$\widehat{\phi}(u) = \widehat{\phi}(0) \prod_{j=1}^{\infty} \widehat{c}\left((M^T)^{-j} u\right), \quad u \in \mathbf{R^d} \tag{1.4}$$

(see [11]). Henceforth, we assume that $\widehat{\phi}(0) = 1$.

An (M, c) refinable function $\phi \in L^2(\mathbf{R^d})$ is *stable* if $\{\phi(x-p)\}_{p\in\mathbf{Z^d}}$ is a Riesz basis of its closed linear span. It is *orthonormal* if $\{\phi(x-p)\}_{p\in\mathbf{Z^d}}$ is an orthonormal basis of its closed linear span.

For an (M, c) refinable function $\phi \in L^2(\mathbf{R})$, define

$$\phi_{\mathrm{au}}(x) := \int_{\mathbf{R^d}} \phi(x-t)\overline{\phi(-t)}\, dt, \quad x \in \mathbf{R^d}, \tag{1.5}$$

and

$$c_{\mathrm{au}}(p) := \sum_{q\in\mathbf{Z^d}} c(p-q)\overline{c(-q)}, \quad p \in \mathbf{Z^d}. \tag{1.6}$$

Then ϕ_{au} is a continuous (M, c_{au}) refinable function. The function ϕ_{au} is called the *autocorrelation* of ϕ and the sequence c_{au} is called the *autocorrelation* of c.

A necessary condition for an (M, c) refinable function ϕ to be orthonormal is that the refinement mask c satisfies the conditions

$$mc_{\mathrm{au}}(Mp) = \delta(p), \quad p \in \mathbf{Z^d}, \tag{1.7}$$

and

$$\sum_{q \in \mathbf{Z^d}} c(q) = 1, \tag{1.8}$$

where $\delta(p) = 1$ for $p = 0$ and $\delta(p) = 0$ otherwise. A sequence c that satisfies (1.7) and (1.8) is called a *conjugate quadrature filter* with respect to the dilation matrix M (M-CQF). Note that (1.8) implies that

$$\sum_{q \in \mathbf{Z^d}} c_{\mathrm{au}}(q) = 1. \tag{1.9}$$

For a dilation matrix M and any finitely supported refinement mask c, we define the (M, c) *subdivision operator* $S_c : \ell^2(\mathbf{Z^d}) \to \ell^2(\mathbf{Z^d})$ by

$$(S_c b)(p) := \sum_{q \in \mathbf{Z^d}} mc(p - Mq) b(q), \quad b \in \ell^1(\mathbf{Z^d}). \tag{1.10}$$

For the case $M = 2I$, this operator has been studied extensively in [1]. The adjoint of $S_{\tilde{c}}$, where $\tilde{c}(p) := \overline{c(-p)},\ p \in \mathbf{Z^d}$, is the (M, c) *transition operator*, which shall be denoted by W_c. Thus the operator $W_c : \ell^2(\mathbf{Z^d}) \to \ell^2(\mathbf{Z^d})$ is defined by

$$(W_c b)(p) = \sum_{q \in \mathbf{Z^d}} mc(Mp - q) b(q), \quad b \in \ell^2(\mathbf{Z^d}). \tag{1.11}$$

We remark that the transition operator $W_{c_{\mathrm{au}}}$ corresponding to the autocorrelation c_{au} of c is called the *wavelet-Galerkin operator* in [18].

Note that if c is conjugate symmetric, i.e., $c = \tilde{c}$, then $S_c = S_{\tilde{c}}$, and W_c is the adjoint of the subdivision operator S_c. For our purposes, we shall restrict the transition operator to the space $\ell^1(\mathbf{Z^d})$. If ϕ is an (M, c) refinable continuous function, where c is finitely supported, then ϕ is compactly supported, and the sequence $\phi_|$ is an eigenvector of W_c in $\ell^1(\mathbf{Z^d})$ of eigenvalue 1. If c is an M-CQF, then $\phi_{c_{\mathrm{au}}|} = \delta$, which is an eigenvector of $W_{c_{\mathrm{au}}}$ of eigenvalue 1.

The definition in (1.11) shows that if the refinement mask c is supported in $[0, N-1]^d$ and if $M = 2I$, then $W_{C_{\mathrm{au}}}$ maps sequences supported on $[-N+1, N-1]^d$ into sequences supported on $[-N+1, N-1]^d$. For a general dilation matrix M, a more detailed discussion in section 4 leads to the fact that $W_{c_{\mathrm{au}}}$ has a finite-dimensional invariant subspace that consists of sequences on a finite set. The operator $W_{c_{\mathrm{au}}}$ (respectively, W_c) restricted to any of its finite-dimensional invariant subspaces will be called a *restricted transition operator*.

The eigenvalues of $W_{C_{\mathrm{au}}}$ hold the key to many important properties of the refinable function, for instance, stability, regularity, and the convergence of the cascade algorithm (see [17], [5], [9], [8], [22], and [23]). The first indication of this role appeared in [17] and [18], where it was shown that for $d = 1$ and $M = (2)$, i.e., in one dimension with dyadic scaling, if ϕ is a $(2, c)$ refinable function, then ϕ is orthonormal if and only if 1 is a simple eigenvalue of the transition operator $W_{c_{\mathrm{au}}}$.

The object of this note is to investigate further the relationship between stability and orthonormality of an (M, c) refinable function on one hand and the eigenvalues of the corresponding transition operator $W_{c_{\mathrm{au}}}$ on the other in the multivariate setting with an arbitrary dilation matrix. In particular, we give a multidimensional extension of the results in [17] on the characterization of the orthonormality of the refinable function. We further show that an (M, c) refinable function ϕ is stable if and only if the transition operator $W_{c_{\mathrm{au}}}$ has a unique eigenvector of eigenvalue 1, whose Fourier transform does not vanish on the torus. This is given in Theorem 2.5, where Theorem 2.3 plays a key role in the proof. Another consequence of Theorem 2.3 is the fact that if the shifts of an (M, c) refinable function ϕ form a Riesz basis, then the sequence $\phi_{c_{\mathrm{au}}|}$ is the unique eigenvector of $W_{c_{\mathrm{au}}}$ corresponding to the eigenvalue 1, and all of the other eigenvalues of $W_{c_{\mathrm{au}}}$ lie inside the unit circle. Section 3 deals with M-CQFs. In particular, it is shown that for an M-CQF, the corresponding (M, c) refinable function belongs to $L^2(\mathbf{R^d})$, and further characterizations of orthonormality are also given. Restricted transition operators are studied in more detail in section 4. It is shown that checking for stability and orthonormality is reduced to checking whether 1 is a simple eigenvalue of a finite-order matrix, which is generated from the refinement mask of ϕ. The general theory is applied to the construction of nonseparable conjugate quadrature filters (M-CQFs) and the corresponding refinable functions from the one-dimensional CQFs of Daubechies.

Another approach to the characterization of stability and orthonormality of a refinable function ϕ with finitely supported refinement mask c makes use of the zero set of the Fourier transform of c. A detailed discussion for the univariate case can be found in [2] and [12]. In one dimension, both approaches characterize the stability and orthonormality of a refinable function ϕ in terms of its refinement mask c using the equivalent characterization of the Fourier transform of ϕ. We prefer the eigenvalue approach for the following reasons. First, as pointed out in [7], for a specified finitely supported mask, it is easier to check for stability and orthonormality of the corresponding refinable function using the eigenvalue characterization. In this case, the problem of checking for stability and orthonormality is reduced to the simple routine of checking whether 1 is a simple eigenvalue of a finite-order matrix. Second, the analysis of the zero set of the Fourier transform of the refinement mask relies on the fact that a univariate polynomial has a finite number of zeros. This no longer holds for multivariate polynomials. However, it is possible to extend the corresponding univariate results to higher dimensions by imposng the condition that a certain multivariate polynomial has a finite number of zeros, as suggested by [11].

It is of particular interest to construct compactly supported wavelets from a compactly supported refinable function and its mask. In the univariate case, with dyadic scaling ($M = (2)$), the construction is simple. For a general integer dilation $M = (m)$, an algorithmic approach in the construction of compactly supported wavelets from a given refinable function ϕ and its refinement mask is given in [16]. The problem of wavelet construction from an (M, c) refinable function ϕ and its refinement mask is much more challenging in higher dimensions. However, in dimensions no greater than 3 and $M = 2I$, a method is given in [20] and [21] under a mild condition on the refinement mask c. In this case, compactly supported wavelets can be constructed based solely on c.

2. Stability and orthonormality of refinable functions.

Let $c : \mathbf{Z^d} \to \mathbf{C}$ be a finitely supported sequence satisfying (1.3) and let M be a dilation matrix. This section studies the relationship between stability and orthonormality of the (M, c)

refinable function ϕ on one hand and the spectral properties of the corresponding transition operator W_c on the other. Recall that the sequence $\phi_|$ is an eigenvector of W_c of eigenvalue 1. We shall first establish a result relating the spectrum of W_c and the nonvanishing of the Fourier transform of $\phi_|$ and then deduce results on stability and orthonormality of ϕ.

LEMMA 2.1. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$. Let ϕ be a continuous (M, c) refinable function and $b \in \ell^1(\mathbf{Z^d})$. Then for any integer $N \geq 1$ and for any $r \in \mathbf{Z^d}$,*

$$\sum_{p\in\mathbf{Z^d}} b(p)\phi\left(r - M^{-N}p\right) = \left(W_c^N(b * \phi_|)\right)(r). \tag{2.1}$$

Proof. The proof is by induction on N. For $N = 1$, applying the refinement equation (1.1) gives

$$\begin{aligned}\sum_{p\in\mathbf{Z^d}} b(p)\phi\left(r - M^{-1}p\right) &= \sum_{p,q\in\mathbf{Z^d}} mb(p)c(q)\phi_|\left(Mr - p - q\right)\\ &= \sum_{q\in\mathbf{Z^d}} mc(q)b * \phi_|(Mr - q)\\ &= \left(W_c(b * \phi_|)\right)(r).\end{aligned}$$

If (2.1) holds for N, then

$$\begin{aligned}\left(W_c^{N+1}(b * \phi_|)\right)(r) &= \sum_{q\in\mathbf{Z^d}} mc(Mr - q)\left(W_c^N(b * \phi_|)\right)\\ &= \sum_{q\in\mathbf{Z^d}} mc(Mr - q)\sum_{p\in\mathbf{Z^d}} b(p)\phi\left(q - M^{-N}p\right).\end{aligned} \tag{2.2}$$

Interchanging the order of summation on the sum in (2.2) followed by a change of index, it can be written as

$$\sum_{p\in\mathbf{Z^d}} b(p)\sum_{k\in\mathbf{Z^d}} mc(k)\phi\left(M(r - M^{-N-1}p) - k\right) = \sum_{p\in\mathbf{Z^d}} b(p)\phi\left(r - M^{-N-1}p\right).$$

The result now follows by induction. □

COROLLARY 2.2. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$ and ϕ is a continuous (M, c) refinable function. Then $\phi_|$ is the unique eigenvector of W_c in $\ell^1(\mathbf{Z^d}) * \phi_|$ of eigenvalue 1.*

Proof. Suppose that $b * \phi_|,\ b \in \ell^1(\mathbf{Z^d})$, is another eigenvector of W_c of eigenvalue 1. Then (2.1) gives

$$\sum_{p\in\mathbf{Z^d}} b(p)\phi\left(r - M^{-N}p\right) = \left(W_c^N(b * \phi_|)\right)(r) = b * \phi_|(r)$$

for all integers $N \geq 0$ and $r \in \mathbf{Z^d}$. Letting $N \to \infty$, we have

$$\left(\sum_{p\in\mathbf{Z^d}} b(p)\right)\phi(r) = b * \phi_|(r),$$

which is equivalent to

$$\left(\sum_{p\in\mathbf{Z^d}} b(p)\right)\widehat{\phi_|}(u) = \widehat{b}(u)\widehat{\phi_|}(u), \quad u\in\mathbf{R^d}.$$

Since $\widehat{\phi_|}$ does not vanish on a set of positive measure, it follows that $\widehat{b}(u)$ is a constant. Equivalently, $b=\alpha\delta$ for some $\alpha\in\mathbf{C}$. Hence $b*\phi_| = \alpha\phi_|$. □

Remark 1. In general, if $\widehat{\phi_|}$ can be factored as

$$\widehat{\phi_|}(u) = \widehat{h}(u)\widehat{g}(u),$$

where $\widehat{g}(u)\neq 0$ for all $u\in\mathbf{R^d}$ and $g\in\ell^1(\mathbf{Z})$, then it follows from Corollary 2.2 and Wiener's theorem that $\phi_|$ is the unique eigenvector of W_c in $h*\ell^1(\mathbf{Z})$ of eigenvalue 1.

THEOREM 2.3. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0)=1$ and ϕ is a continuous (M,c) refinable function. If*

$$\widehat{\phi_|}(u) := \sum_{p\in\mathbf{Z^d}} \phi(p)e^{-ipu} \neq 0, \quad u\in\mathbf{R^d}, \tag{2.3}$$

then $\phi_|$ is the unique eigenvector of W_c in $\ell^1(\mathbf{Z^d})$ of eigenvalue 1, *and all of the other eigenvalues of W_c lie inside the unit circle.*

Further, 1 *is a simple eigenvalue of any restricted W_c.*

Proof. Suppose that (2.3) holds, and let

$$1/\widehat{\phi_|}(u) =: \sum_{p\in\mathbf{Z^d}} w(p)e^{-ipu}, \quad u\in\mathbf{R^d}.$$

Since $1/\widehat{\phi_|}(u)$ is smooth, it follows that $w\in\ell^1(\mathbf{Z^d})$. Hence $\ell^1(\mathbf{Z^d})*\phi_| = \ell^1(\mathbf{Z^d})$, and Corollary 2.2 implies that $\phi_|$ is the unique eigenvector of W_c in $\ell^1(\mathbf{Z})$ of eigenvalue 1.

Now let $\lambda\neq 1$ be an eigenvalue of W_c and let $v\in\ell^1(\mathbf{Z^d})$ be the corresponding eigenvector. Equation (2.1) gives

$$\lambda^N v(r) = (W_c^N v)(r) = (W_c^N v*w*\phi_|)(r) = \sum_{p\in\mathbf{Z}} (v*w)(p)\phi\left(r-M^{-N}p\right). \tag{2.4}$$

The limit as $N\to\infty$ of the sum on the right of (2.4) exists and is equal to

$$\left(\sum_p v*w(p)\right)\phi(r), \quad r\in\mathbf{Z^d}.$$

Therefore, if $\lambda\neq 1$, then necessarily $|\lambda|<1$. Further, $\sum_p v*w(p)=0$.

If 1 is not a simple eigenvalue of a restricted transition operator W_c, then it must be a degenerate eigenvalue with only one eigenvector, say b. In this case, there exists a vector b_1 such that $W_c b_1 = b_1 + b$, which implies that $W_c^N b_1 = b_1 + Nb$ for all integers $N\geq 1$. Again, (2.4) gives

$$b_1(r) + Nb(r) = (W_c^N b_1)(r) = \sum_{p\in\mathbf{Z}} (b_1*w)(p)\phi\left(r-M^{-N}p\right)$$

for all $N \geq 1$, which is impossible. □

A function $\phi \in L^2(\mathbf{R^d})$ is stable if $\{\phi(\cdot - p)\}_{p \in \mathbf{Z^d}}$ is a Riesz basis of its closed linear span. Recall that $\{\phi(\cdot - p)\}_{p \in \mathbf{Z}}$ is a Riesz basis if and only if there exist constants $0 < C_1 \leq C_2 < \infty$ such that

$$C_1 \leq \sum_{q \in \mathbf{Z^d}} |\hat{\phi}(u + 2\pi q)|^2 \leq C_2 \quad \text{for almost all } u \in \mathbf{R^d}. \tag{2.5}$$

If $\phi_{\mathrm{au}|} \in \ell^1(\mathbf{Z^d})$, the Poisson summation formula leads to the characterization that ϕ is stable if and only if

$$C_1 \leq \sum_p \phi_{\mathrm{au}}(p) e^{-ipu} \leq C_2, \quad u \in \mathbf{R^d}. \tag{2.6}$$

Corollary 2.4. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$ and that ϕ is a stable (M, c) refinable function. Then $\phi_{\mathrm{au}|}$ is the unique eigenvector in $\ell^1(\mathbf{Z^d})$ of $W_{c_{\mathrm{au}}}$ corresponding to the eigenvalue 1, and all of the other eigenvalues of $W_{c_{\mathrm{au}}}$ lie inside the unit circle.*

Corollary 2.4 follows directly from Theorem 2.3 and equation (2.6). In one dimension with dyadic scaling, Cohen and Daubechies [3] proved that for a stable $(2, c)$ refinable function ϕ, the eigenvalues of the corresponding transition operator $W_{c_{\mathrm{au}}}$ restricted to its invariant subspace of finite sequences with zero mean lie inside the unit circle. Their result was extended to higher dimensions with dilation matrix $M = 2I$ by Long and Chen [14]. We note that in one dimension, the result of Cohen and Daubechies was also improved upon by Hervé [10].

The following theorem gives a characterization of the stability of an (M, c) refinable function. A similar result in one dimension with dilation $M = (2)$ was obtained in [5].

Theorem 2.5. *Suppose that c is a finitely supported sequence that satisfies $\widehat{c}(0) = 1$. An (M, c) refinable function ϕ is stable if and only if $W_{c_{\mathrm{au}}}$ has a unique eigenvector of eigenvalue 1 whose Fourier transform does not vanish.*

Further, 1 is a simple eigenvalue of any restricted $W_{c_{\mathrm{au}}}$.

Proof. If ϕ is stable, then condition (2.3) of Proposition 2.3 is satisfied for ϕ_{au}. Hence $W_{c_{\mathrm{au}}}$ has a unique eigenvector $\phi_{\mathrm{au}|}$ of eigenvalue 1 that has a nonvanishing Fourier transform.

Conversely, since $\phi_{\mathrm{au}|} \in \ell^1(\mathbf{Z^d})$ is an eigenvector of $W_{c_{\mathrm{au}}}$ with eigenvalue 1 and since such an eigenvector is unique and has a nonvanishing Fourier transform, it follows from (2.6) that the (M, c) refinable function ϕ is stable. □

An (M, c) refinable function $\phi \in L^2(\mathbf{R^d})$ is *interpolatory* if ϕ is continuous and satisfies

$$\phi(p) = \delta(p), \quad p \in \mathbf{Z^d}. \tag{2.7}$$

Theorem 2.6. *Suppose that c is a finitely supported sequence that satisfies $\widehat{c}(0) = 1$. A necessary and sufficient condition for a continuous (M, c) refinable function ϕ to be interpolatory is that the sequence δ is the unique eigenvector of W_c of eigenvalue 1.*

Further, 1 is a simple eigenvalue of any restricted $W_{c_{\mathrm{au}}}$.

Proof. Since ϕ is (M, c) refinable, the corresponding sequence $\phi_|$ is an eigenvector of W_c in $\ell^1(\mathbf{Z^d})$ with eigenvalue 1. If δ is the unique eigenvector of eigenvalue 1, then

$$\phi(p) = \delta(p), \quad p \in \mathbf{Z^d},$$

i.e., ϕ is interpolatory.

Conversely, if ϕ is interpolatory, then obviously $\phi_| = \delta$ is an eigenvector of W_c in $\ell^1(\mathbf{Z^d})$ with eigenvalue 1. Since

$$\sum_{p\in\mathbf{Z^d}} \phi(p)e^{-ipu} = 1 \quad \text{for all } u \in \mathbf{R^d}$$

does not vanish, by Theorem 2.3, δ is the unique eigenvector of eigenvalue 1. □

Clearly, $\phi \in L^2(\mathbf{R^d})$ is an orthonormal (M,c) refinable function if and only if ϕ_{au} is an interpolatory (M, c_{au}) refinable function.

Corollary 2.7. *Suppose that c is a finitely supported sequence that satisfies $\widehat{c}(0) = 1$. An (M,c) refinable function ϕ is orthonormal if and only if the sequence δ is the unique eigenvector of $W_{c_{\text{au}}}$ of eigenvalue* 1.

Further, 1 *is a simple eigenvalue of any restricted $W_{c_{\text{au}}}$.*

Combining this corollary with Theorem 2.5, we have the following proposition.

Proposition 2.8. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$ and that $\phi \in L^2(\mathbf{R^d})$ is (M,c) refinable. Then the following are equivalent:*

(i) *ϕ is orthonormal*;

(ii) *c is an M-CQF and ϕ is stable*;

(iii) *c is an M-CQF and $\widehat{\phi_{\text{au}|}}(u) \neq 0, u \in \mathbf{R^d}$.*

If c is an M-CQF, Corollary 2.7 says that the simplicity of the eigenvalue 1 of any restricted transition operator $W_{c_{\text{au}}}$ is equivalent to the orthonormality of the refinable function ϕ. On the other hand, Theorem 2.5 states that the stability of the refinable function ϕ is equivalent to the simplicity of the eigenvalue 1 of any restricted transition operator $W_{c_{\text{au}}}$ and the nonvanishing of the Fourier transform of the corresponding eigenvector. It will be shown in an example below that the simplicity of the eigenvalue 1 of a restricted $W_{c_{\text{au}}}$ does not imply the existence of an eigenvector with nonvanishing Fourier transform. This shows that the conditions in Theorem 2.5 are not superfluous. In fact, it will be interesting to know whether the simplicity of the eigenvalue 1 of a restricted transition operator together with the additional condition

$$\sum_k |\widehat{c}(u + (M^T)^{-1}2\pi\gamma_k)|^2 > 0 \tag{2.8}$$

would imply the nonvanishing of the Fourier transform of the corresponding eigenvector.

The following examples show that the simplicity of eigenvalue 1 together with (2.8) do not imply the nonvanishing of the Fourier transform of the corresponding eigenvector. In particular, they show that the converse of Theorem 2.3 is false even under the assumption in (2.8).

Example 1. Let c be the sequence

$$\left\{\ldots, 0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0, \ldots\right\},$$

where $c(0) = 1/4$, $c(1) = 1/2, \ldots$. It is straightforward to check that the sequence $a = \{\ldots, 0, 1, 2, 2, 2, 1, 0, \ldots\}$, where $a(0) = 1$, $a(1) = 2, \ldots$, is the unique eigenvector for W_c of eigenvalue 1 and that the Fourier transform of a is $(1 + e^{iu})(1 + e^{iu} + e^{i2u})$, which vanishes at -1.

We shall show that there is a compactly supported continuous $(2,c)$ refinable function ϕ such that

$$\phi(n) = a(n), \quad n \in \mathbf{Z}, \tag{2.9}$$

and the Fourier transform $\widehat{c}(u)$ of the mask c satisfies

$$|\widehat{c}(u)|^2 + |\widehat{c}(u+\pi)|^2 > 0 \quad \text{for all } u \in \mathbf{R}. \tag{2.10}$$

The Fourier transform of c can be written as

$$\widehat{c}(u) = \widehat{b}(u)(1 - e^{iu} + e^{i2u}) = \widehat{b}(u)(e^{iu} + \omega)(e^{iu} + \omega^2), \tag{2.11}$$

where

$$b(u) = \frac{1}{4}(1 + e^{iu})^3$$

is the Fourier transform of the mask

$$b = \left\{ \dots, 0, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, 0, \dots \right\}$$

with $b(0) = 1/4,\ b(1) = 3/4, \dots$, and $\omega \neq 1$ is a cube root of unity. All of the roots of $\widehat{c}$ lie on the unit circle, and they are precisely $-1, -1, -1, -\omega$, and $-\omega^2$. Since no root is the negative of another root, $|\widehat{c}(u)|^2 + |\widehat{c}(u+\pi)|^2 > 0$ for all real u. Thus the mask c satisfies (2.10).

Note that the sequence b is exactly the mask for the $(2,b)$ refinable quadratic B-spline g obtained by convolving the characteristic function $\chi_{[0,1)}$ with itself three times. Let

$$\phi(x) = g(x) + g(x-1) + g(x-2), \quad x \in \mathbf{R}. \tag{2.12}$$

Then the Fourier transform of ϕ is

$$\widehat{\phi}(u) = \widehat{g}(u)(e^{iu} - \omega)(e^{iu} - \omega^2) = \widehat{g}(u)(1 + e^{iu} + e^{i2u}). \tag{2.13}$$

Since the set of roots of $(1-z)^3(1+z+z^2)$ is closed under the mapping $z \to z^2$, $\phi(x)$ is $(2,a)$ refinable by Theorem 2.1 of [15]. The function ϕ is not stable since ω and ω^2 are zeros of $1 + z + z^2$. With a suitable normalization of g, we have

$$\phi_| = \{\dots, 0, 1, 2, 2, 2, 1, 0, \dots\},$$

which is (2.9).

Example 2. Let c and ϕ be as in Example 1. Then

$$c_{\text{au}} = \{\dots, 0, 1, 4, 6, 6, 9, 12, 9, 6, 6, 4, 1, 0, \dots\}.$$

The sequence $\phi_{\text{au}|}$ is an eigenvector of $W_{c_{\text{au}}}$ of eigenvalue 1. The 11×11 linear system satisfied by the eigenvectors for $W_{c_{\text{au}}}$ corresponding to the eigenvalue 1 has rank 10. (Here $W_{c_{\text{au}}}$ is restricted to sequences supported on $[-5,5]$, which form an invariant subspace containing all finitely supported eigenvectors.) Hence $\phi_|$ is the unique eigenvector of $W_{c_{\text{au}}}$ of eigenvalue 1.

In summary, this example gives a $(2,c)$ refinable function which is not stable, but $W_{c_{\text{au}}}$ has a simple eigenvalue 1 and the mask c satisfies

$$|\widehat{c}(u)|^2 + |\widehat{c}(u+\pi)|^2 > 0 \quad \text{for all } u.$$

3. Multivariate CQF. Let M be an integer dilation matrix and $c : \mathbf{Z^d} \to \mathbf{C}$ be a finitely supported sequence that satisfies (1.3). Hence there is a unique compactly supported (M, c) refinable distribution ϕ, normalized so that $\widehat{\phi}(0) = 1$. Assuming that c is an M-CQF, we are interested in knowing when $\phi \in L^2(\mathbf{R^d})$ and obtaining further characterizations of orthonormality.

We first consider the cascade algorithm for the computation of the compactly supported (M, c) refinable distribution. Let $\phi^{(0)}$ be the indicator function of any fundamental region for $\mathbf{Z^d} \subset \mathbf{R^d}$. (Thus $\widehat{\phi^{(0)}}$ is continuous at 0 and $\widehat{\phi^{(0)}}(0) = 1$.) Starting from $\phi^{(0)}$, define a sequence of functions $\phi^{(n)}$ by

$$\phi^{(n)}(x) := \sum_{p \in Z^d} mc(p)\phi^{(n-1)}(Mx - p), \quad n = 1, 2, \ldots. \tag{3.1}$$

Then

$$\widehat{\phi^{(n)}}(u) = \widehat{\phi^{(0)}}\left((M^T)^{-n}u\right) \prod_{j=1}^{n} \left(\widehat{c}\left((M^T)^{-j}u\right)\right), \quad u \in \mathbf{R^d}. \tag{3.2}$$

The sequence $\widehat{\phi^{(n)}} \to \widehat{\phi}$ uniformly on compact subsets as $n \to \infty$, where

$$\widehat{\phi}(u) = \prod_{j=1}^{\infty} \widehat{c}\left((M^T)^{-j}u\right), \quad u \in \mathbf{R^d}, \tag{3.3}$$

since c satisfies (1.3). Further, $\widehat{\phi}$ is continuous at the origin and $\widehat{\phi}(0) = 1$.

It is clear that $\phi^{(n)} \in L^2(\mathbf{R^d})$ and is compactly supported. Next, we prove that $\|\phi^{(n)}\| = 1$ for all $n = 0, 1, \ldots$. Note that if c is an M-CQF, then (1.7) and (1.11) imply that δ is an eigenvector of $W_{c_{\mathrm{au}}}$ of eigenvalue 1. Since $\phi^{(0)}_{\mathrm{au}\,|} = \delta$ and $\phi^{(n)}_{\mathrm{au}\,|} = W_{c_{\mathrm{au}}} \phi^{(n-1)}_{\mathrm{au}\,|}$, we have $\phi^{(n)}_{\mathrm{au}\,|} = \delta$. Hence $\|\phi^{(n)}\| = 1$ for all $n = 0, 1, \ldots$.

PROPOSITION 3.1. *If c is an M-CQF, then $\phi^{(n)}$ defined by* (3.1) *converges weakly to the (M, c) refinable function $\phi \in L^2(\mathbf{R^d})$.*

If, in addition, $\|\phi\| = 1$, then $\phi^{(n)}$ converges strongly in $L^2(\mathbf{R^d})$ to ϕ.

Proof. First, note that if c is an M-CQF, then $||\phi^{(n)}|| = 1$ for all $n \geq 0$. Therefore, $\{\phi^{(n)}\}$ has a subsequence which converges weakly to $\varphi \in L^2(\mathbf{R^d})$. Since weak convergence is stronger than convergence in distribution, we have $\varphi = \phi$, and hence $\phi \in L^2(\mathbf{R^d})$.

Next, we show that the sequence $\phi^{(n)}$ itself converges weakly to ϕ. If $\phi^{(n)}$ does not converge weakly to ϕ, then there exists a subsequence $\phi^{(n_i)}$ which converges weakly to a function in $L^2(\mathbf{R^d})$ other than ϕ. This contradicts the fact that $\phi^{(n_i)}$ converges to ϕ in distribution.

In addition, if $\|\phi\| = 1$, then $\|\phi^{(n)}\| \to \|\phi\|$. With this, weak convergence of $\phi^{(n)} \to \phi$ implies strong convergence. □

Remark 2. For a general finitely supported mask c, a similar proof shows that the corresponding (M, c) refinable distribution ϕ belongs to $L^2(\mathbf{R^d})$ if the ℓ^2 operator norm $\|W_{c_{\mathrm{au}}}\| \leq 1$.

Since ϕ is compactly supported, if $\phi \in L^2(\mathbf{R^d})$, then $\phi \in L^1(\mathbf{R^d})$. Hence $\widehat{\phi}(u) \to 0$ as $|u| \to \infty$. If ϕ is refinable, then for any $p \in 2\pi\mathbf{Z^d}/\{\mathbf{0}\}$,

$$\widehat{\phi}((M^T)^n p) = \widehat{\phi}(p) \prod_{j=1}^{n} \widehat{c}(((M^T)^{n-j})p) = \widehat{\phi}(p).$$

Letting $n \to \infty$ implies that $\widehat{\phi}(p) = 0$ for $p \in 2\pi\mathbf{Z^d}/\{\mathbf{0}\}$. This means that ϕ satisfies the Strang–Fix condition of order 1. By the Poisson summation formula, this condition is equivalent to the shifts of ϕ forming a partition of unity, i.e.,

$$\sum_{p\in\mathbf{Z^d}} \phi(x-p) = 1, \quad x \in \mathbf{R^d}. \tag{3.4}$$

Proposition 3.2. *Let M be a dilation matrix, c be an M-CQF, and ϕ be the unique (M,c) refinable function normalized such that $\widehat{\phi}(0) = 1$. The following are equivalent:*

(i) *δ is the unique eigenvector of $W_{c_{\mathrm{au}}}$ of eigenvalue* 1;

(ii) *the shifts of ϕ are orthonormal*;

(iii) *the shifts of ϕ are orthogonal.*

Proof. The equivalence of (i) and (ii) is given in Corollary 2.7. We need only to show that (iii) implies (ii). If (iii) holds, multiplying both sides of (3.4) by ϕ and integrating term by term, the orthogonality of the shifts of ϕ and the fact that $\int \phi(x) = 1$ give $\|\phi\|^2 = 1$. □

As a consequence of Propositions 3.1 and 3.2, the cascade algorithm converges strongly if the corresponding (M,c) refinable function ϕ is orthonormal, a result which coincides with the well-known fact that the stability of an (M,c) refinable L^2 function implies strong convergence of the cascade algorithm.

4. Restricted transition operators. We now discuss how to restrict the transition operator to a finite-dimensional subspace. For a dilation matrix M and a finitely supported refinement mask c, a subset $D \subset \mathbf{Z^d}$ is called an *invariant support set* for the transition operator W_c if the following are satisfied:

(i) D is finite;

(ii) for all sequences b with support in D, the support of $W_c b$ is also in D; and

(iii) the support of every finitely supported eigenvector of W_c that corresponds to a nonzero eigenvalue is contained in D.

Such a finite invariant support set D for W_c always exists. To construct D, choose a vector norm $\|\cdot\|$ on $\mathbf{R^d}$ and a number $0 < \alpha < 1$ such that for all $x \in \mathbf{R^d}$,

$$||M^{-1}x|| \le (1-\alpha)||x||.$$

This is possible because the spectral radius $\rho(M^{-1}) < 1$. Now choose

$$r \ge r_{\text{min}} := \frac{(1-\alpha)}{\alpha} \max_{c(p)\neq 0} \|p\|$$

and define

$$D_r = \{p \in \mathbf{Z^d} : ||\mathbf{p}|| \le \mathrm{r}\}.$$

Clearly, D_r is an invariant support set for W_c. Further, if a sequence b is supported in D_s with $s > r$, then $W_c b$ is supported in D_t, where $t = \alpha r + (1-\alpha)s$. Therefore, D_r contains the support of every compactly supported eigenvector of W_c. Further, any compactly supported eigenvector of W_c is also an eigenvector of the restricted operator $W_c|_{\ell(D_r)}$, where $\ell(D_r)$ is the space of all sequences supported on D_r. One may construct the disk D_r to be arbitrary close to the minimal size by using a vector norm so that the corresponding operator norm $\|M^{-1}\|$ is sufficiently closed to $\rho(M^{-1})$ and by choosing $r = r_{\text{min}}$.

For the case where $M = 2I$ with refinement mask c supported in $[0, N-1]^d$, the support of ϕ is in $[0, N-1]^d$ and that of ϕ_{au} is contained in $[-N+1, N-1]^d$. A mimimal invariant support set for the corresponding transition operator $W_{c_{\mathrm{au}}}$ is $[-N+1, N-1]^d$.

For a dilation matrix M and a finitely supported mask c, let Ω be an invariant support set of $W_{c_{\mathrm{au}}}$, and $\ell(\Omega)$ be the space of all sequences supported on Ω. Then the transition operator $W_{c_{\mathrm{au}}}$ restricted to $\ell(\Omega)$ is represented by the matrix

$$A := (m\ c_{\mathrm{au}}(Mp - q))_{p,q\in\Omega}\ , \tag{4.1}$$

and $\phi_{\mathrm{au}|}$ is an eigenvector of A of eigenvalue 1.

Theorem 4.1. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$ and that $\phi \in L^2(\mathbf{R^d})$ is the compactly supported (M, c) refinable function. Then ϕ is stable if and only if*

(i) *there is an eigenvector corresponding to the eigenvalue 1 of matrix A defined by (4.1) whose Fourier transform does not vanish, and*

(ii) *1 is a simple eigenvalue of A.*

Proof. Conditions (i) and (ii) together with the fact that $\phi_{\mathrm{au}|}$ is an eigenvector of A of eigenvalue 1 imply that $\phi_{\mathrm{au}|}$ has a nonvanishing Fourier transform. Hence conditions (i) and (ii) imply that ϕ is stable.

On the other hand, if ϕ is stable, then $\phi_{\mathrm{au}|}$ is an eigenvector of A of eigenvalue 1 whose Fourier transform does not vanish; hence condition (i) holds. To show condition (ii), assume that 1 is not a simple eigenvalue of A. Then there exists an eigenvector a of A of eigenvalue 1, and a is not a scalar multiple of the eigenvector $\phi_{\mathrm{au}|}$. Since the transition operator $W_{c_{\mathrm{au}}}$ maps $\ell(\Omega)$ into $\ell(\Omega)$, the vector a is an eigenvector in $\ell^1(\mathbf{Z})$ of the transition operator $W_{c_{\mathrm{au}}}$ of eigenvalue 1. This contradicts Theorem 2.5. □

A similar argument using Theorem 2.6 and Corollary 2.7, respectively, leads to the following results.

Proposition 4.2. *Let c be a finitely supported sequence that satisfies $\widehat{c}(0) = 1$, ϕ be the (M, c) refinable function, and D be an invariant support set of W_c, and suppose that ϕ is continuous. Then ϕ is interpolatory if and only if the sequence δ is a unique eigenvector of the matrix*

$$C := (m\ c(Mp - q))_{p,q\in D}$$

of simple eigenvalue 1.

Proposition 4.3. *Suppose that c is a finitely supported sequence satisfying $\widehat{c}(0) = 1$ and that ϕ is the (M, c) refinable function in $L^2(\mathbf{R^d})$. Then ϕ is orthonormal if and only if the sequence δ is a unique eigenvector of the matrix A defined by (4.1) of simple eigenvalue 1.*

This proposition shows that the problem of checking whether ϕ has orthonormal shifts simply amounts to checking whether 1 is a simple eigenvalue of the matrix A. Similarly, checking whether ϕ is stable reduces to checking whether 1 is a simple eigenvalue of the matrix A and whether the Fourier transform of the corresponding eigenvector vanishes on the torus.

In the case where $d = 2, 3$ and $M = 2I$, if ϕ has orthonormal shifts and the refinement mask c satisfies

$$\overline{\widehat{c}(u)} = e^{ip_0\cdot y}\widehat{c}(u)$$

for some $p_0 \in \mathbf{Z^d}$, then it was shown in [20] and [21] that compactly supported orthonormal wavelets can easily be constructed from c and ϕ. Interested readers should consult [20] and [21] for details.

5. Construction of admissible refinement masks. This section constructs three 2×2 dilation matrices M with $|\det(M)| = 2$ and an infinite family of two-dimensional finitely supported masks c and shows that the corresponding (M, c) refinable functions ϕ are orthonormal. The refinable functions ϕ are constructed so that the set of points satisfying the condition $\{u : \widehat{\phi}(u) \neq 0\}$ contains a connected open set that contains a fundamental domain for $2\pi\mathbf{Z^2} \subset \mathbf{R^2}$. This implies that $(\widehat{\phi_{\text{au}|}})(u) > 0$; hence by Corollary 2.7, ϕ has orthonormal shifts.

Up to a similarity transformation and multiplication by matrices that represent reflection about the origin and reflection about the x_1 axis, there are only three distinct 2×2 integer dilation matrices whose determinant equals 2 or -2. They are

$$M_1 = \begin{pmatrix} 0 & -2 \\ 1 & 0 \end{pmatrix}, \quad \text{which has eigenvalues } \pm i\sqrt{2},$$

$$M_2 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \text{which has eigenvalues } 1 \pm i,$$

and

$$M_3 = \begin{pmatrix} 0 & -2 \\ 1 & 1 \end{pmatrix}, \quad \text{which has eigenvalues } \frac{(1 \pm i\sqrt{7})}{2}.$$

We shall now construct refinement masks $c : \mathbf{Z}^2 \to \mathbf{C}$ which generate nonseparable orthonormal refinable functions and wavelets for dilation matrices M_1, M_2, and M_3.

For a given one-dimensional sequence b, we define the *induced two-dimensional mask* $c : \mathbf{Z^2} \to \mathbf{C}$ by

$$c\begin{pmatrix} m \\ n \end{pmatrix} = b(m)\delta(n), \quad (m, n)^T \in \mathbf{Z^2}. \tag{5.1}$$

LEMMA 5.1. *Let $Z_{\widehat{b}}$ denote the set of real zeros of $\widehat{b}$. Then the zero set $Z_{\widehat{c}}$ of the Fourier transform of the induced mask c is given by*

$$Z_{\widehat{c}} = \left\{(u_1, u_2)^T : u_1 \in Z_{\widehat{b}},\ u_2 \in R\right\}. \tag{5.2}$$

Proof. From (5.1),

$$\widehat{c}((u_1, u_2)^T) = \widehat{b}(u_1), \quad (u_1, u_2)^T \in \mathbf{R^2},$$

which gives (5.2). □

LEMMA 5.2. *Let b and c be as above. Let M be a 2×2 dilation matrix, and let ϕ be the unique (M, c) refinable distribution. Then the zero set $Z_{\widehat{\phi}}$ of the Fourier transform of ϕ satisfies*

$$Z_{\widehat{\phi}} = \bigcup_{j \geq 1} (M^T)^j Z_{\widehat{c}}. \tag{5.3}$$

Proof. The assertion follows from the infinite-product representation of $\widehat{\phi}$. □

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be the lines $x_1 = \pi$ and $x_1 = -\pi$ in $\mathbf{R^2}$, respectively. For any 2×2 dilation matrix M and the corresponding (M, c) refinable function ϕ, the set K

is defined as the closure of the largest connected subset of $\mathbf{R^2}$ containing the origin and consisting of points where the Fourier transform $\widehat{\phi}$ is nonzero.

Lemma 5.3. *Let b be a one-dimensional mask and let c be the induced two-dimensional mask. Suppose that the zero set of $\hat{b}$ satisfies*

$$Z_{\widehat{b}} = \{(2n+1)\pi : n \in \mathbf{Z}\}. \tag{5.4}$$

Then K is bounded by a subset of lines $(M^T)^k \mathcal{L}_j,\ k \geq 1,\ j = 1, 2.$

Proof. The assertion follows from Lemmas 5.1 and 5.2. □

Remark 3. The refinement masks b used by Daubechies in [6] to construct orthonormal refinable functions of one variable are CQFs, and the zero set of $\widehat{b}$ satisfies condition (5.4).

For the three dilation matrices $M_n,\ n = 1, 2, 3$, the set K can be computed explicitly if the zero set of $\hat{b}$ satisfies condition (5.4). In each case, the set K is a polygon whose vertices are the columns of the matrix V_n, where

$$V_1 = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

$$V_2 = \begin{pmatrix} 0 & 1 & \frac{3}{2} & 0 & -1 & -\frac{3}{2} \\ \\ 1 & 1 & \frac{1}{2} & -1 & -1 & -\frac{1}{2} \end{pmatrix},$$

$$V_3 = \begin{pmatrix} \frac{2}{3} & \frac{4}{3} & -\frac{2}{3} & -\frac{4}{3} \\ \\ \frac{5}{3} & \frac{1}{3} & -\frac{5}{3} & -\frac{1}{3} \end{pmatrix}.$$

Theorem 5.4. *Suppose that b is a one-dimensional mask satisfying the zero condition* (5.4)*, and let c be the two-dimensional induced mask. Then for $M = M_n$, $n = 1, 2, 3$, the symbol $\widehat{\phi_{\mathrm{au}|}}(u)$ is positive. Further, if b is a CQF, then c is also an M-CQF, and the corresponding (M, c) refinable function is orthonormal.*

Proof. It is straightforward to check that for each of the three dilation matrices $M_n,\ n = 1, 2, 3$, the interior of K contains the closure of a fundamental domain for $\mathbf{Z^2} \subset \mathbf{R^2}$. Therefore, the symbol $\widehat{\phi_{\mathrm{au}|}}(u) > 0$.

If b is a CQF, then c is an M-CQF because the intersection of $M\mathbf{Z^2}$ with the lattice points on the x_1-axis is exactly the set of even integers. Since the symbol is positive, by Corollary 2.7, the integer shifts of ϕ are orthonormal. □

We note that the results of Theorem 5.4 for the dilation matrix M_2 have been obtained by Cohen and Daubechies [4].

Remark 4. The fact that $\widehat{\phi_{\mathrm{au}|}}(u) > 0$ implies orthonormality was first proved in [19]. However, the proof in that paper was based on the Lebesque dominated convergence theorem and special properties of scaling tiles and was quite complicated. In [19], the refinable functions produced above were also constructed and mesh plots of some of these functions were produced. However, the report did not examine the zero set of $\widehat{\phi}$ and therefore did not actually prove that the refinable functions constructed had orthonormal shifts.

Remark 5. For each of the dilation matrices $M_n, n = 1, 2, 3$ and the mask induced by the Daubechies length-4 coefficients b, we computed an invariant support set for the transition operator $W_{c_{\mathrm{au}}}$, and we computed the eigenvalues of the corresponding restricted transition operator. There are 49, 63, and 39 nonzero eigenvalues (counted with multiplicity) corresponding to M_1, M_2, and M_3, respectively, and the eigenvalue

1 is simple in all cases. All of the eigenvalues have modulus ≤ 1 and several—but not all—of these eigenvalues are negative-integer powers of the eigenvalues of the corresponding dilation matrix. This is significant because the degree of smoothness of the refinable function implies the existence of a finite number of such eigenvalues. The corresponding eigenvectors can be constructed from the derivatives of the refinable function in the directions of the eigenvectors of the dilation matrix. The degree of smoothness of the refinable function also implies the existence of a continuous spectrum for the unrestricted transition operator that includes a continuous family of eigenvectors constructed from fractional derivative and integral operators. The discrete spectrum of the transition operator can easily be shown to coincide with the spectrum of the restricted transition operator. The significance of discrete eigenvalues that do not correspond to negative integer powers of the eigenvalues of the dilation matrices will be discussed in a subsequent paper.

Remark 6. The existence of negative-integer powers of the eigenvalues of the dilation matrix in the spectrum of $W_{c_{\mathrm{au}}}$ is not sufficient for regularity of the corresponding refinable function. Indeed, the (M_2, c) refinable function ϕ constructed from the Daubechies length-4 sequence is not continuous [24, Example 5.2], and none of the longer filters leads to C^1 solutions [4, Theorem 4.2].

Remark 7. Let V_0 be the closed shift-invariant subspace generated by ϕ and let

$$V_k := \{f(M^k\cdot) : f \in V_0\}.$$

Then $\{V_k\}$ forms a multiresolution analysis of $L^2(\mathbf{R}^\mathbf{d})$ by Remark 2.6 of [13]. We further remark that the construction of the corresponding wavelet for ϕ and c is straightforward since $|\det M| = 2$.

Acknowledgment. We thank an anonymous referee for Remark 6 and for some relevant references.

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision,* Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[2] A. Cohen, *Ondelettes, analyses, multirésolution et traitement numérique du signal*, Ph.D. thesis, Université Paris Dauphine, Paris, 1990.

[3] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their subband coding scheme,* Duke Math. J., 68 (1992), pp. 313–335.

[4] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases* 1, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[5] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal basis of compactly supported wavelets,* Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelet,* Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[7] I. Daubechies, *Ten Lectures on Wavelets*, CBMS Regional Conference Series in Applied Mathematics, Vol. 61, SIAM, Philadephia, 1992.

[8] T. Eirola, *Sobolev characterization of solutions of dilation equations,* SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[9] P. N. Heller and R. O. Wells, Jr., *The spectral theory of multiresolution operators and applications,* Technical report, AWARE, Bedford, MA, 1993.

[10] L. Hervé, *Construction et regularite des fonctions d' echelle,* SIAM J. Math. Anal., 26 (1995), pp. 1361–1385.

[11] T. Hogan, *Stability and Independence of the Shifts of a Multivariate Refinable Function*, 1995, preprint.

[12] R.-Q. Jia and J. Z. Wang, *Stability and linear independence associated with wavelet decompositions*, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.

[13] R.-Q. Jia and Z. Shen, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc. (2), 37 (1994), pp. 271–300.

[14] R. Long and D. Chen, *Biorthogonal wavelet bases on* $\mathbf{R^d}$, Appl. Comput. Harmonic Anal., 2 (1995), pp. 230–242.

[15] W. Lawton, S. L. Lee, and Z. Shen, *Complete characterization of refinable splines*, Adv. Comput. Math., 3 (1995), pp. 137–145.

[16] W. Lawton, S. L. Lee, and Z. Shen, *An algorithm for matrix extension and wavelet construction*, Math. Comp., 65 (1996), pp. 723–737.

[17] W. Lawton, *Necessary and sufficient conditions for constructing orthonormal wavelets*, J. Math. Phys., 32 (1991), pp. 52–61.

[18] W. Lawton, *Multilevel properties of the wavelet-Galerkin operator,* J. Math. Phys., 32 (1991), pp. 1440–1443.

[19] W. Lawton and H. Resnikoff, *Multidimensional wavelet bases*, Technical report, AWARE, Bedford, MA, 1991.

[20] S. D. Riemenschneider and Z. Shen, *Box splines, cardinal series, and wavelets*, in Approximation Theory and Functional Analysis, C. K. Chui, ed., Academic Press, New York, 1991, pp. 133–149.

[21] S. D. Riemenschneider and Z. Shen, *Wavelets and pre-wavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.

[22] G. Strang, *Eigenvalues of* $(\downarrow 2)H$ *and Convergence of the Cascade Algorithm,* preprint, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[23] L. F. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets,* SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[24] L. F. Villemoes, *Continuity of nonseparable quincux wavelets,* Appl. Comput. Harmonic Anal., 1 (1994), pp. 180–187.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1015–1027, September 1997

BOLTZMANN EQUATION WITH INFINITE ENERGY: RENORMALIZED SOLUTIONS AND DISTRIBUTIONAL SOLUTIONS FOR SMALL INITIAL DATA AND INITIAL DATA CLOSE TO A MAXWELLIAN*

S. MISCHLER† AND B. PERTHAME‡

Abstract. We prove new existence results for the Boltzmann equation with an initial data with infinite energy. In the framework of renormalized solutions we assume $(|x|^\alpha + |x-v|^2)\, f_0 \in L^1$ instead of $(|x|^2+|v|^2)\, f_0 \in L^1$, and we show new a priori estimates. In the framework of distributional solutions we treat small initial data compared to a Maxwellian of the type $\exp(-|x-v|^2/2)$. We also treat initial data close enough to such a Maxwellian. Hence, our theory does not require that the initial data decrease in both variables x and v.

Key words. Boltzmann equation, moments lemma, renormalized solution, distributional solution, small initial data, initial data close to a Maxwellian, dispersive effects

AMS subject classifications. 35Q21, 76P05, 35D05, 82C40

PII. S0036141096298102

1. Introduction.

This paper is devoted to the existence proof of solutions to the Boltzmann equation in the case of initial data with infinite energy. We prove global existence of renormalized solutions and distributional solutions either for small initial data compared to the local Maxwellian $\exp(-|x-v|^2/2)$ or for initial data close to that particular local Maxwellian. Hence, we generalize the classical theories to initial data which do not decay in all the directions of the phase space. They may have infinite mass and energy.

More precisely, we consider the Boltzmann equation which describes the statistical evolution of a moderately rarefied gas. In this model, the gas is described by the kinetic density $f(t,x,v) \geq 0$ of particles which at time $t \in [0,+\infty[$, at the point $x \in \mathbb{R}^N$, move with velocity $v \in \mathbb{R}^N$, where N is an integer ≥ 1. This kinetic density satisfies the Boltzmann equation

$$(1.1)\quad \begin{cases} \dfrac{\partial f}{\partial t} + v\cdot\nabla_x f = Q(f,f) = Q^+(f,f) - Q^-(f,f) \qquad \text{on} \quad]0,+\infty[\times\mathbb{R}^N\times\mathbb{R}^N, \\ f(0,x,v) = f_0(x,v) \qquad\quad \text{on} \quad \mathbb{R}^N\times\mathbb{R}^N. \end{cases}$$

We refer to Cercignani [C], Cercignani, Illner, and Pulvirenti [C,I,P], DiPerna and Lions [DP,L1] and to their references for a detailed presentation of the physical meaning and the notion of renormalized solution of such an equation; we will recall this definition in section 3.

Here, Q is a quadratic collision operator, acting only on velocities, and is defined by

$$(1.2)\qquad Q^+(\varphi,\varphi)(v) = \int_{v_\star\in\mathbb{R}^N}\int_{\omega\in S^{N-1}} \varphi'\,\varphi'_\star\, B(v-v_\star,\omega)\,d\omega dv_\star,$$

*Received by the editors January 15, 1996; accepted for publication (in revised form) April 29, 1996.

http://www.siam.org/journals/sima/28-5/29810.html

†Département de Mathématiques, Université de Versailles-Saint-Quentin, B timent Fermat, 45, avenue des États-Unis, 78055 Versailles Cedex, France (mischler@math.uvsq.fr).

‡Ecole Normale Supérieure, DMI, 45, rue d'Ulm, 75230 Paris Cédex 05, France (perthame@ann.jussieu.fr).

$$Q^-(\varphi,\varphi) = \varphi\, L(\varphi), \tag{1.3}$$

where

$$L(\varphi) = A *_v \varphi \quad \text{ and } \quad A(z) = \int_{S^{N-1}} B(z,\omega)\, d\omega. \tag{1.4}$$

As usual, we denote, in order to shorten notation, $\varphi = \varphi(v)$, $\varphi_\star = \varphi(v_\star)$, $\varphi' = \varphi(v')$, and $\varphi'_\star = \varphi(v'_\star)$. The post-collisional velocities v' and $v'_\star$ are obtained from the pre-collisional velocities v and $v_\star$ and the unit vector ω thanks to

$$\begin{cases} v' = v - \langle v - v_\star, \omega\rangle \omega, \\ v'_\star = v_\star + \langle v - v_\star, \omega\rangle \omega. \end{cases} \tag{1.5}$$

Here and everywhere below, we denote indifferently by $a{\cdot}b$ or $\langle a, b\rangle$ the usual scalar product of $a, b \in \mathbb{R}^N$. Equations (1.5) are just a parametrization of the conservation of the impulsion and the kinetic energy during the collisions

$$\begin{cases} v' + v'_\star = v + v_\star, \\ |v'|^2 + |v'_\star|^2 = |v|^2 + |v_\star|^2. \end{cases} \tag{1.6}$$

The collision kernel B that enters the bilinear operator Q is a given function on $\mathbb{R}^N \times S^{N-1}$. We will always assume that

$$B \geq 0\,, \quad B(z,\omega) \text{ depends only on } |z| \text{ and } |\langle z,\omega\rangle| \tag{1.7}$$

and the so-called Grad [Gr] angular cut-off assumption

$$B \in L^1_{\rm loc}(\mathbb{R}^N \times S^{N-1}). \tag{1.8}$$

Renormalized solutions for the Boltzmann equation have been introduced and developed by DiPerna and Lions [DP,L1], [DP,L2], [L1], [L2], [L3], [L4]. They have proved stability results and the existence of global solutions for initial data with finite mass, energy, and entropy

$$\iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)\Big(1 + |x|^\alpha + |v|^2 + |\log f_0(x,v)|\Big)\, dvdx < +\infty, \tag{1.9}$$

for some $\alpha > 0$. In [L4], existence is extended to initial data which are bounded perturbations of particular solutions, such as pure Maxwellian.

In this work we do not assume bounded energy anymore. Due to dispersion effects (see Perthame [P1]), the assumption (1.9) can be replaced, for instance, by

$$\Gamma_\alpha = \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)\Big(1 + |x|^\alpha + |x-v|^2 + |\log f_0(x,v)|\Big)\, dvdx < +\infty, \tag{1.10}$$

for some $\alpha > 0$. Using the fact that $\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} f(t,x,v)\, |x-(1+t)\, v|^2\, dvdx$ is independant of t, we prove velocity moment bounds which give us enough a priori estimates to apply the stability result of renormalized solutions. Notice that for the Vlasov–Poisson system also, one can build solutions with infinite energy. It is enough to assume $f_0 \in L^1 \cap L^\infty$ and $|x|^2 f_0 \in L^1$ (see [P1]). Notice also that our assumption on f_0 implies that $(1+|x|^\alpha+|v|^\alpha+|\log f_0(x,v)|)\, f_0 \in L^1$. But the existence of renormalized solutions under this only assumption is open. Also, the time decay for solutions built with only the assumption (1.10) is an open question. Under the assumption (1.9), an answer is given in [P1].

On the other hand, existence, uniqueness, and time decay of the global distributional solution have been studied by many authors. The recent theory was initiated by Illner and Shinbrot [I,S] (see also [K,S]). Much progress has been made [Ba,D,G], [B,P,T], [B,T], [H], [Pa] and most general assumptions are due to [T]. They deal with small initial data with respect to a reference function $\varphi(v)\,h(x)$. The functions φ and h can have Maxwellian decay or, in the latest works, polynomial decay, which allow initial data with infinite mass; see [T]. Other existence results have been obtained for initial data close to a local Maxwellian. These theories rely on dispersive effects in the whole space.

Using the same idea as above, namely, that $|x-(1+t)\,v|^2$ solves the free transport equation and a is collisional invariant, we construct global upper and lower Maxwellian solutions of the Boltzmann equation. By the standard maximum principle this proves existence results. The new fact here is the decay assumption only on the direction $x-v$ instead of both directions x and v. We can deal with two situations: small initial data compared to the Maxwellian $\exp(-|x-v|^2/2)$ for general kernel Q or initial data close to such a Maxwellian for Maxwellian molecules. These solutions have infinite energy; nevertheless, we can prove that energy becomes locally finite: particles with high energy go away very fast despite the collisions.

The outline of the paper is the following. In the second section we establish new estimates for a classical solution of the Boltzmann equation. We prove velocity moment bounds and a local energy bound in some cases when the initial data does not satisfy energy bound. We use appropriate multiplicators in the spirit of Perthame [P2] and Lions and Perthame [L,P]. In the third section we use these estimates in order to prove the existence of renormalized solutions to the Boltzmann equation in an infinite energy case. In the fourth section we prove the existence of the global classical solution for small initial data with infinite mass and energy. In the fifth section we deal with initial data close to a local Maxwellian.

2. A priori estimates. In this section we prove new estimates for a classical solution f to the Boltzmann equation. In order to rigorously establish our result, we assume that f is of class C^1 and has compact support in space and velocity variables for every fixed time t. We assume that the collision kernel B just satisfies (1.7) and (1.8). It is classical to prove by change of variables and thanks to the symmetry property (1.7) that, for all test functions ψ in $L^\infty_{\text{loc}}(\mathbb{R}^N)$, we have the following equality:

(2.1)
$$\begin{aligned}&\int_{\mathbb{R}^N} Q(f,f)\psi dv\\ &= \frac{1}{4}\int_{v\in\mathbb{R}^N}\int_{v_\star\in\mathbb{R}^N}\int_{\omega\in S^{N-1}} B(v-v_\star,\omega)(f'\,f'_\star - f\,f_\star)(\psi+\psi_\star-\psi'-\psi'_\star)\,d\omega dv_\star dv.\end{aligned}$$

This implies that 1, v, and $|v|^2$ are collisional invariant; indeed, taking $\psi=1$, $(v_i)_{1\le i\le N}$, $|v|^2$ in (2.1), and using equation (1.6), we have

$$\int_{\mathbb{R}^N} Q(f,f)\begin{pmatrix}1\\ v\\ |v|^2\end{pmatrix} dv = 0 \quad \forall\, t,\, x\in[0,+\infty)\times\mathbb{R}^N. \tag{2.2}$$

Lemma 1. *For all time $t \geq 0$, the solution f satisfies*

$$\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)|x-(t+1)\,v|^2\,dvdx = \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)|x-v|^2\,dvdx. \tag{2.3}$$

Proof of Lemma 1. We just multiply the Boltzmann equation (1.1) by $|x-(t+1)\,v|^2$:

$$\left(\frac{\partial}{\partial t}+v\cdot\nabla_x\right)(f\,|x-(t+1)\,v|^2) = Q(f,f)\,|x-(t+1)\,v|^2.$$

Then, we integrate the previous equation in velocity and space variables, and because $|x-(t+1)\,v|^2$ is a collisional invariant, the right-hand side vanishes and we obtain the result. □

Lemma 2. *Let α be a given real number in $(0,2)$; then the solution f satisfies for every $t \geq 0$:*

$$\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)(1+|x|^2)^{\alpha/2}\,dvdx \leq e^{\alpha\,t}\iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)\Big((1+|x|^2)^{\alpha/2}+|x-v|^2\Big)\,dvdx. \tag{2.4}$$

Proof of Lemma 2. We multiply the Boltzmann equation (1.1) by $(1+|x|^2)^{\alpha/2}$ and because $(1+|x|^2)^{\alpha/2}$ is a collisional invariant, we get after integration

$$\frac{d}{dt}\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)(1+|x|^2)^{\alpha/2}\,dvdx = \alpha\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\,v\cdot x\,(1+|x|^2)^{\alpha/2-1}\,dvdx.$$

Next, using

$$2|v\cdot x| \leq |v|^2+|x|^2 \leq \left(1+\frac{1}{(1+t)^2}\right)|x|^2+\frac{1}{(1+t)^2}|x-(t+1)\,v|^2,$$

we get

$$\begin{aligned}&\frac{d}{dt}\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)(1+|x|^2)^{\alpha/2}\,dvdx\\ &\quad\leq \alpha\left(\frac{1}{2}+\frac{1}{2(1+t)^2}\right)\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\frac{|x|^2}{1+|x|^2}(1+|x|^2)^{\alpha/2}\,dvdx\\ &\qquad\qquad+\frac{\alpha}{2(1+t)^2}\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\frac{|x-(t+1)\,v|^2}{(1+|x|^2)^{1-\alpha/2}}\,dvdx\\ &\quad\leq \alpha\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)(1+|x|^2)^{\alpha/2}\,dvdx+\alpha\iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)|x-v|^2\,dvdx.\end{aligned}$$

We conclude the proof thanks to the Gronwall lemma. □

Lemma 3. *Let $\alpha \in (0,2)$; if Γ_α is finite (see (1.10)), there exists a constant $C_T = C(T,\alpha,\Gamma_\alpha)$ such that*

$$\sup_{[0,T]}\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\Big(1+|x|^\alpha+|v|^\alpha+|\log f(t,x,v)|\Big)\,dvdx \leq C_T, \tag{2.5}$$

$$\int_0^T\iint_{\mathbb{R}^N\times\mathbb{R}^N} e(f)\,dvdxdt \leq C_T, \tag{2.6}$$

where

$$e(f)=\iint_{\mathbb{R}^N\times S^{N-1}} B(v-v_\star,\omega)(f'\,f'_\star-f\,f_\star)\log\frac{f'\,f'_\star}{f\,f_\star}\,d\omega dv_\star. \tag{2.7}$$

Proof of Lemma 3. We compute

$$\begin{aligned}&\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)|v|^\alpha\,dvdx \le \iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\frac{1}{(1+t)^\alpha}\Big(|x|+|x-(t+1)\,v|\Big)^\alpha\,dvdx\\ &\le C\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\Big((1+|x|^2)^{\alpha/2}+|x-(t+1)\,v|^2\Big)\,dvdx,\end{aligned}$$

which is bounded by $C_T\,\Gamma_\alpha$ thanks to Lemmas 1 and 2.

The entropy estimate is classically deduced from the bound (2.6) and we skip it (see [DP,L1] for the case $\alpha=2$). □

LEMMA 4. *For every T, there exists a constant C_T such that the solution f satisfies*

$$\int_0^T\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\frac{|v|^2}{(1+|x|^2)^{3/2}}\,dvdxdt\le C_T\,\Gamma_1. \tag{2.8}$$

Proof of Lemma 4. We multiply (1.1) by $\dfrac{x\cdot v}{(1+|x|^2)^{1/2}}$, and we compute

$$v\cdot\nabla_x\Big(\frac{x\cdot v}{(1+|x|^2)^{1/2}}\Big)=\frac{|v|^2}{(1+|x|^2)^{3/2}}+\frac{|v|^2|x|^2(1-\frac{x}{|x|}\cdot\frac{v}{|v|})}{(1+|x|^2)^{3/2}}.$$

We remark that the second right-hand term is nonnegative, and thus we get

$$\begin{aligned}&\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(T,x,v)\frac{x\cdot v}{(1+|x|^2)^{1/2}}\,dvdx-\iint_{\mathbb{R}^N\times\mathbb{R}^N} f_0(x,v)\frac{x\cdot v}{(1+|x|^2)^{1/2}}\,dvdx\\ &\ge\int_0^T\iint_{\mathbb{R}^N\times\mathbb{R}^N} f(t,x,v)\frac{|v|^2}{(1+|x|^2)^{3/2}}\,dvdxdt.\end{aligned}$$

This ends the proof because the two left-hand side terms are bounded by $C_T\,\Gamma_1$ thanks to Lemma 3.

3. Renormalized solution. In this section we show how the previous estimates can be used to prove an existence result of a renormalized solution to the Boltzmann equation for some initial data with infinite energy. DiPerna and Lions [DP,L1] have introduced the concept of renormalized solution. A function $f\in C([0,+\infty[;L^1(\mathbb{R}^N\times\mathbb{R}^N))$ is a renormalized solution of the Boltzmann equation if

$$\frac{Q^\pm(f,f)}{1+f}\in L^1_{\mathrm{loc}}([0,+\infty[\times\mathbb{R}^N\times\mathbb{R}^N), \tag{3.1}$$

$$\frac{\partial}{\partial t}\log(1+f)+v\cdot\nabla_x\log(1+f)=\frac{Q(f,f)}{1+f}\quad\text{in}\quad\mathcal{D}'([0,\infty[\times\mathbb{R}^N\times\mathbb{R}^N). \tag{3.2}$$

They have proved a stability result that we mimic here in order to prove our existence result.

In this section we assume, following [DP,L1], that for a given $\bar{\alpha} \in (0,1)$ or $\bar{\alpha} = 2$

$$\frac{1}{1+|z|^{\bar{\alpha}}} \int_{B_R} A(v-z)\,dv \underset{|z|\to+\infty}{\longrightarrow} 0 \quad \forall R \in]0,+\infty[, \tag{3.3}$$

where $B_R = \{v \in \mathbb{R}^N; |v| < R\}$.

THEOREM 1. *Let f_0 be an initial data such that in* (1.10), $\Gamma_\alpha(f_0) < +\infty$ *for a given $\alpha \in (0,1]$ and B such that* (1.7), (1.8), *and* (3.3) *hold with $\bar{\alpha} = \alpha$ if $\alpha < 1$ and with $\bar{\alpha} = 2$ if $\alpha = 1$. Then there exists a global renormalized solution f of the Boltzmann equation with initial data f_0, and* (2.5), (2.6), *and* (2.8) *hold.*

Proof of Theorem 1. We split the proof in two steps.

Step 1. A regularized problem.

For $\lambda > 1$ we define the following cross-section:

$$B_\lambda(v,v_\star,\omega) = \frac{\lambda}{\text{meas}(S^{N-1})} \wedge B(v-v_\star,\omega)\cdot \mathcal{X}_\lambda(v,v_\star,\omega) \tag{3.4}$$

$$\text{where } \mathcal{X}_\lambda(v,v_\star,\omega) = \begin{cases} 1 & \text{if } v, v_\star, v' \text{ and } v'_\star \text{ belong to } B(0,\lambda), \\ 0 & \text{elsewhere,} \end{cases}$$

so that

$$0 \le A_\lambda(v,v_\star) = \int_{S^{N-1}} B_\lambda(v,v_\star,\omega)\,d\omega \le \lambda. \tag{3.5}$$

We set $q_\lambda(\varphi) = r_\lambda(\int_{\mathbb{R}^N} \varphi\,dv)$, where r_λ is a smooth function such that $r_\lambda = 1$ if $t < \frac{\lambda}{2}$, $r_\lambda \le \frac{\lambda}{t}$ for every t.

Let $(\rho_\epsilon(x,v))_{\epsilon>0}$ be a sequence of mollifiers, with $\text{supp}\,\rho_\epsilon \subset B(0,1)$ for all ϵ, and let

$$\phi_\lambda = \left(f_0(x,v)\, 1_{\left\{|v|<\frac{\lambda}{2}\right\}} 1_{\left\{|x|<\frac{\lambda}{2}\right\}}\right) \star \rho_{\frac{1}{\lambda}}. \tag{3.6}$$

We are looking for a solution to the regularized problem

$$\begin{cases} \dfrac{\partial f_\lambda}{\partial t} + v\cdot\nabla_x f_\lambda = q_\lambda(f_\lambda)\, Q_\lambda(f_\lambda,f_\lambda), \\ f_\lambda(0,x,v) = \phi_\lambda(x,v), \end{cases} \tag{3.7}$$

where Q_λ is defined from B_λ in the same way that Q from B in (1.2) to (1.4). We remark that this collisional term has a linear growth at infinity.

In order to prove the existence of a solution we set

$$\mathcal{A} = \{\varphi \ge 0, \quad \varphi \in L^\infty([0,T]\times\mathbb{R}^N\times\mathbb{R}^N), \quad \text{supp}\,\varphi(t,.) \subset B_{(1+t)\lambda}\times B_\lambda\},$$

which is a complete set for the uniform norm. Next, we define the map $\varphi \in \mathcal{A} \longmapsto \Lambda(\varphi) = \psi$ as the solution of

$$\begin{cases} \dot{\psi}^\# + \lambda\psi^\# = q_\lambda(\varphi)^\#\, Q^+_\lambda(\varphi,\varphi)^\# + \big(\lambda - q_\lambda(\varphi)^\#\, L_\lambda(\varphi)^\#\big)\varphi^\#, \\ \psi^\#(0,x,v) = \phi_\lambda(x,v), \end{cases} \tag{3.8}$$

where we denote $R^{\#}(t,x,v) = R(t,x+t\,v,v)$. It is obvious that Λ maps $\mathcal{A}$ into $\mathcal{A}$ and is Lipschitz continuous for the uniform norm. Then the Cauchy–Lipschitz theorem and the Gronwall lemma imply the existence of a unique global solution f_λ of the equation (3.7), which falls in $\mathcal{A}$ and is smooth.

Step 2. Passing to the limit.

Lemmas 3 and 4 applied to the sequence of regularized solution f_λ imply that f_λ satisfies the bounds (2.5), (2.6), and (2.8) uniformly in λ. Then the stability results for renormalized solution of DiPerna and Lions [DP,L1], [DP,L2], [L3] adapt here and allow us to conclude. □

4. Distributional solution for small initial data. We use the same idea in the theory of distributional solution, and more precisely for solutions satisfying $Q(f,f) \in L^\infty$. The proof we present here is adapted from [L2], and we refer to [B,P,T] for a general presentation of the method as well as a very sharp assumption, which, however, is different from ours.

Theorem 2. *Let f_0 satisfy, for some a, b, $C_0 > 0$ and x_0, $v_0 \in \mathbb{R}^N$,*

$$0 \le f_0(x,v) \le \hat{f}_0(x,v) := \frac{C_0}{6} \exp\left(-\frac{1}{2}\left|\frac{x-x_0}{a} - \frac{v-v_0}{b}\right|^2\right). \tag{4.1}$$

We assume that $A \in L^{p'}(\mathbb{R}^N)$ for some $p' \in (\frac{N}{N-1}, +\infty]$, and we have

$$C_0\, b^{\frac{N}{p}-1}\, a < \kappa := \frac{N-p}{p\,\|A\|_{p'}} \left(\frac{p}{2\pi}\right)^{\frac{N}{2p}}. \tag{4.2}$$

Then there exists a solution $f \in L^\infty((0,+\infty)\times\mathbb{R}^N_x\times\mathbb{R}^N_v)$ to the Boltzmann equation (1.1) *in $\mathcal{D}'([0,\infty[\,\times\mathbb{R}^N\times\mathbb{R}^N)$, such that*

$$0 \le f(t,x,v) \le \frac{C(t)}{6} \exp\left(-\frac{1}{2}\left|\frac{x-x_0-v\,t}{a} - \frac{v-v_0}{b}\right|^2\right), \tag{4.3}$$

where $C(t)(\ge 0)$ is uniformly bounded; more precisely,

$$\frac{1}{C(t)} = \frac{1}{C_0} - \frac{b^{\frac{N}{p}-1}\,a}{\kappa} + \frac{a}{\kappa}\left(\frac{1}{b}+\frac{t}{a}\right)^{1-\frac{N}{p}}. \tag{4.4}$$

Also, $Q^\pm(f,f) \in L^\infty((0,T)\times\mathbb{R}^N\times\mathbb{R}^N)$.

Proof of Theorem 2. We set

$$\hat{f} = C(t)\exp\left(-\frac{1}{2}\left|\frac{x-x_0-v\,t}{a} - \frac{v-v_0}{b}\right|^2\right),$$

where $C(t) \ge 0$ will be determined in such a way that we have

$$\dot{C}C^{-1}\hat{f} \ge Q^+(\hat{f},\hat{f}) \quad \text{on} \quad [0,\infty[\,\times\mathbb{R}^N\times\mathbb{R}^N. \tag{4.5}$$

We remark that $\hat{f}$ is, for all $x \in \mathbb{R}^N$ and $t \ge 0$, a Maxwellian in v, and thus $Q^+(\hat{f},\hat{f}) \equiv Q^-(\hat{f},\hat{f}) = \hat{f}\cdot A \star_v \hat{f}$.

Therefore, (4.5) holds if we have

$$\dot{C} \ge C^2 \sup_{(x,v)\in\mathbb{R}^{2N}} \left\{ A \star_v \exp\left(-\frac{1}{2}\left|\frac{x-x_0-v\,t}{a} - \frac{v-v_0}{b}\right|^2\right)\right\}.$$

Next we observe that we have on $\mathbb{R}^{2N}$, thanks to a Hölder inequality with p and p' such that $\frac{1}{p}+\frac{1}{p'}=1$ (which implies $\frac{N}{p}>1$),

$$\begin{aligned}
&A\star_v \exp\left(-\frac{1}{2}\left|\frac{x-x_0-v\,t}{a}-\frac{v-v_0}{b}\right|^2\right)\\
&\qquad\qquad\le \|A\|_{p'}\left(\int_{\mathbb{R}^N}\exp\left(-p\left|\frac{x-x_0}{a}+\frac{v_0}{b}-\left(\frac{1}{b}+\frac{t}{a}\right)v\right|^2\right)dv\right)^{1/p}\\
&\qquad\qquad\le \|A\|_{p'}\left(\frac{2\pi}{p}\right)^{\frac{N}{2p}}\left(\frac{1}{b}+\frac{t}{a}\right)^{-\frac{N}{p}} =: \lambda(t).
\end{aligned}$$

Therefore, we choose $C\in C^1([0,\infty[)$ such that $C(0)=C_0$ and solves the ordinary differential equation

$$\dot{C}=\frac{C^2}{\kappa}\frac{1}{\frac{N}{p}-1}\left(\frac{1}{b}+\frac{t}{a}\right)^{-\frac{N}{p}}. \tag{4.6}$$

The solution to (4.6) is given by (4.4) and is a global solution if the condition (4.2) holds.

In such a way, we have constructed, dividing $\hat{f}$ by 6, an upper solution $\hat{f}$ of the Boltzmann equation in the following sense:

$$\begin{cases}\dfrac{\partial}{\partial t}\hat{f}+v\cdot\nabla_x\hat{f}\ge 6\,\hat{\lambda}(t)\,\hat{f}\ge 6\,Q^{\pm}(\hat{f},\hat{f}) \quad \text{in} \quad \mathcal{D}'([0,\infty[\times\mathbb{R}^N\times\mathbb{R}^N),\\ \hat{f}(t=0,.)=\hat{f}_0,\end{cases} \tag{4.7}$$

where $\hat{\lambda}(t)=C(t)\,\lambda(t)$.

In order to prove the existence result, we define a suitable function space where we will do a classical Banach fixed point theorem. Let $\mathcal{B}$ be the space

$$\mathcal{B}=\left\{\varphi\in L^\infty([0,T]\times\mathbb{R}^N\times\mathbb{R}^N);\ 0\le\varphi(t,x,v)\le\hat{f}(t,x,v)\right\}, \tag{4.8}$$

and we define the map $\varphi\in\mathcal{B}\longmapsto\Lambda(\varphi)=\psi$ by

$$\begin{cases}\dfrac{\partial}{\partial t}\psi+v\cdot\nabla_x\psi+\hat{\lambda}(t)\,\psi=Q^+(\varphi,\varphi)+(\hat{\lambda}(t)-L(\varphi))\varphi,\\ \psi(t=0,.)=f_0.\end{cases} \tag{4.9}$$

The norm defined in $\mathcal{B}$ is the following:

$$\|\varphi\|=\operatorname*{ess\,sup}_{t\in[0,T]\,x\in\mathbb{R}^N\,v\in\mathbb{R}^N}\frac{|\varphi(t,x,v)|}{\hat{f}(t,x,v)}.$$

The function space $\mathcal{B}$ is, with the previous norm, a Banach space. The global existence and uniqueness of the solution of the Boltzmann equation is assured if the following conditions hold:

$$\forall\varphi\in\mathcal{B};\quad \Lambda\varphi\in\mathcal{B}, \tag{4.10}$$

$$\forall\varphi_1,\varphi_2\in\mathcal{B};\quad \|\Lambda\varphi_2-\Lambda\varphi_1\|\le\frac{5}{6}\,\|\varphi_2-\varphi_1\|. \tag{4.11}$$

First, we show (4.10). For all $\varphi \in \mathcal{B}$ we have

$$\begin{aligned}\frac{\partial}{\partial t}\psi + v \cdot \nabla_x \psi + \hat{\lambda}(t)\,\psi &= Q^+(\varphi,\varphi) + (\hat{\lambda}(t) - L(\varphi))\varphi,\\ &\le Q^+(\hat{f},\hat{f}) + \hat{\lambda}(t)\,\hat{f},\\ &\le \frac{\partial}{\partial t}\hat{f} + v \cdot \nabla_x \hat{f} + \hat{\lambda}(t)\hat{f}.\end{aligned}$$

Thus $\psi \le \hat{f}$. On the other hand, $\psi \ge 0$ is clear since the source term in (4.9) is nonnegative thanks to the definition of $\hat{\lambda}(t)$.

We pass to (4.11). For all $\varphi_1, \varphi_2 \in \mathcal{B}$ the following equality holds:

$$\begin{aligned}\left(\frac{\partial}{\partial t} + v \cdot \nabla_x + \hat{\lambda}(t)\right)(\psi_2 - \psi_1) &= Q^+(\varphi_2, \varphi_2 - \varphi_1) + Q^+(\varphi_2 - \varphi_1, \varphi_1)\\ &\quad + \hat{\lambda}(t)(\varphi_2 - \varphi_1) - L(\varphi_2)(\varphi_2 - \varphi_1) - (L(\varphi_2) - L(\varphi_1))\varphi_1.\end{aligned}\tag{4.12}$$

Denoting by $\mathcal{C}(t,x,v)$ the right-hand side term, equation (4.13) writes

$$(\psi_2 - \psi_1)^{\#}(t,x,v) = \int_0^t \mathcal{C}^{\#}(s,x,v)e - \int_s^t \hat{\lambda}(\sigma)\,d\sigma\,ds.$$

Next, multiplying and dividing all terms of the right-hand side by $\hat{f}^{\#}$ and using the definition of the norm and (4.7) yield

$$\begin{aligned}|\psi_2 - \psi_1|^{\#}(t,x,v) &\le (2\,\|\varphi_1\| + 2\,\|\varphi_2\| + 1)\,\|\varphi_1 - \varphi_2\| \int_0^t \hat{\lambda}(s)\,\hat{f}^{\#}(s,x,v)\,ds\\ &\le \frac{5}{6}\,\|\varphi_1 - \varphi_2\|\,\hat{f}^{\#}(t,x,v),\end{aligned}$$

and thus (4.11) holds. □

Remarks 1. We can examine in which sense the initial data is small. Of course, for a given Maxwellian profile (a, b, x_0, and v_0 fixed) f_0 has to be small enough for the uniform norm with respect to $\exp\bigl(-\frac{1}{2}\bigl|\frac{x-x_0}{a} - \frac{v-v_0}{b}\bigr|^2\bigr)$, but depending on a and b, $\|f_0\|_\infty$, can be as large as we wish.

Next, at the macroscopic level, the global mass is always infinite. Also, the macroscopic functions associated with the Maxwellian distribution are

$$\begin{aligned}\rho_0(t,x) &= \bar{\rho}_0 = (2\,\pi)^{\frac{N}{2}}\,b^N\,\frac{C_0}{6},\\ u_0(t,x) &= v_0 + \frac{b}{a}\,(x - x_0),\\ T_0(t,x) &= b^2.\end{aligned}$$

Thus, we see that we can take the uniform norm of f_0, ρ_0, and T_0 arbitrarily large if we choose $\frac{a}{b}$ small enough such that (4.2) holds. In that case the quantity

$$\sup_{v \in \mathbb{R}^N} \int_{x \in \mathbb{R}^N} f_0(x,v)\,dx \le \bar{\rho}_0\,\bigl(\frac{a}{b}\bigr)^N$$

is small; in particular, f_0 is small in $L^1(B_R \times B_R)$ for all fixed $R > 0$.

Remarks 2. We also recover the classical time decay. From (4.2) and (4.4), $C(t)$ is uniformly bounded (by some constant C_∞), and an explicit computation of $\hat{f}$ gives

$$\rho(t,x) \le \frac{C_\infty}{(\frac{1}{b}+\frac{t}{a})^N},$$
$$\rho(t,x)\,(|u(t,x)|^2+T(t,x)) \le \frac{C_\infty}{(\frac{1}{b}+\frac{t}{a})^{N+2}}.$$

5. Distributional solution for an initial data close to a local Maxwellian.

In this section we show how the idea developed in the previous section can be used to prove the existence of a global distributional solution for an initial data close to a local Maxwellian. We consider the simple case when the reference Maxwellian is $\xi_0(x,v)=\exp(-\frac{|x-v|^2}{2})$ and the cross-section B is of Maxwellian type; precisely, we assume that

$$B(z,\omega)=B\left(\left|\frac{z}{|z|}\cdot\omega\right|\right) \qquad \text{and} \qquad A(z)\equiv(2\,\pi)^{-\frac{N}{2}}. \tag{5.1}$$

Again, this gives a variant of classical results [B,P,T], [K,S] that requires global decay in both directions x and v.

THEOREM 3. *Let f_0 be an initial data which satisfies, for some ϵ, $C_0>0$,*

$$(1-\epsilon)\,C_0\,\xi_0(x,v) \le f_0(x,v) \le (1+\epsilon)\,C_0\,\xi_0(x,v). \tag{5.2}$$

If ϵ is small enough, there exists a distributional solution $f\in L^\infty((0,+\infty)\times\mathbb{R}^N_x\times\mathbb{R}^N_v)$ of the Boltzmann equation (1.1) *in the sense of distribution. Moreover, $Q^\pm(f,f)\in L^\infty((0,+\infty)\times\mathbb{R}^N\times\mathbb{R}^N)$.*

Proof of Theorem 3. *Step* 1. We first build two functions $g_0(t,x,v)=c(t)\,\xi(t,x,v)$ and $G_0(t,x,v)=C(t)\,\xi(t,x,v)$, where $\xi(t,x,v)=\exp(-\frac{|x-v\,(1+t)|^2}{2})$ and $0\le c(t)\le C(t)$ will be determined in such a way that we have

$$\begin{cases} \dfrac{\partial}{\partial t}g_0+v\cdot\nabla_x g_0+L(G_0)\,g_0=Q^+(g_0,g_0),\\ \dfrac{\partial}{\partial t}G_0+v\cdot\nabla_x G_0+L(g_0)\,G_0=Q^+(G_0,G_0),\\ g_0(t=0,.)=(1-\epsilon)\,C_0\,\xi_0, \quad G_0(t=0,.)=(1+\epsilon)\,C_0\,\xi_0. \end{cases} \tag{5.3}$$

To do so, since ξ is a Maxwellian in v and thanks to assumption (5.1), we easily compute

$$Q^+(\xi,\xi)=\xi\,L(\xi)=\frac{\xi}{(1+t)^N}. \tag{5.4}$$

Hence, the equations (5.3) are equivalent to the system of ODEs

$$\begin{cases} \dot{c}+\dfrac{C\,c}{(1+t)^N}=\dfrac{c^2}{(1+t)^N},\\ \dot{C}+\dfrac{c\,C}{(1+t)^N}=\dfrac{C^2}{(1+t)^N},\\ c(0)=(1-\epsilon)\,C_0, \quad C(0)=(1+\epsilon)\,C_0. \end{cases} \tag{5.5}$$

In order to prove the existence of global nonnegative solutions c and C for (5.5), we set $y = C - c$ and $z = C + c$, and (5.5) is equivalent to

$$\begin{cases} \dot{y} = \dfrac{y\,z}{(1+t)^N}, \quad \dot{z} = \dfrac{y^2}{(1+t)^N}, \\ y(0) = 2\,\epsilon\,C_0, \quad z(0) = 2\,C_0. \end{cases} \tag{5.6}$$

It is enough to prove an a priori bound for y. Eliminating z, the system (5.6) reduces to

$$\dot{y} = \frac{y}{(1+t)^N}\Big(2\,C_0 + \int_0^t \frac{y^2(s)\,ds}{(1+s)^N}\Big). \tag{5.7}$$

First, we remark that since $\dot{y} \geq 0$ we have

$$\dot{y} \leq 2\,C_0 \frac{y}{(1+t)^N} + \frac{y^3}{(1+s)^{2N-1}}. \tag{5.8}$$

Next, we consider the simple equation

$$\dot{Y} = (2\,C_0 + 1)\frac{Y}{(1+t)^N}, \quad Y(0) = y(0), \tag{5.9}$$

which has a global solution satisfying $Y(t) \leq 1$ if

$$\epsilon \leq \frac{1}{2\,C_0}\exp\big(2\,C_0\,(1-N)\big). \tag{5.10}$$

Now, by a comparison principle, we get $y(t) \leq Y(t)$ for all $t \geq 0$. This ends the proof of the existence of g_0 and G_0.

Step 2. Following the classical Kaniel and Shinbrot iterative scheme [K,S], we define the sequences $(g_n)_{n\in\mathbb{N}}$ and $(G_n)_{n\in\mathbb{N}}$ as

$$\begin{cases} \dfrac{\partial}{\partial t} g_n + v\cdot\nabla_x g_n + L(G_{n-1})\,g_n = Q^+(g_{n-1}, g_{n-1}), \\ \dfrac{\partial}{\partial t} G_n + v\cdot\nabla_x G_n + L(g_{n-1})\,G_n = Q^+(G_{n-1}, G_{n-1}), \\ g_n(t=0,.) = G_n(t=0,.) = f_0. \end{cases} \tag{5.11}$$

As in [K,S] we obtain, by a comparison principle, that g_n, G_n satisfy

$$0 \leq g_0(t) \leq g_1(t) \leq \cdots \leq g_n(t) \leq \cdots \leq G_n(t) \leq \cdots \leq G_1(t) \leq G_0(t). \tag{5.12}$$

Therefore, g_n and G_n are monotone sequences and converge pointwise to limits denoted by $\bar{g}$ and $\bar{G}$. We may pass to the limit, in the distributional sense, in (5.11), and we obtain

$$\begin{cases} \dfrac{\partial}{\partial t} \bar{g} + v\cdot\nabla_x \bar{g} + L(\bar{G})\,\bar{g} = Q^+(\bar{g}, \bar{g}), \\ \dfrac{\partial}{\partial t} \bar{G} + v\cdot\nabla_x \bar{G} + L(\bar{g})\,\bar{G} = Q^+(\bar{G}, \bar{G}), \\ \bar{g}(t=0,.) = \bar{G}(t=0,.) = f_0. \end{cases} \tag{5.13}$$

It remains to show that $\bar{G} = \bar{g}$. We remark that thanks to (5.12) we already know that $\bar{G} \geq \bar{g}$. To prove the other inequality, notice that

$$\begin{cases} \dfrac{\partial}{\partial t}(\bar{G}-\bar{g}) + v\cdot\nabla_x(\bar{G}-\bar{g}) + (\bar{G}-\bar{g})\,L(\bar{G}) \\ \qquad\qquad = Q^+(\bar{G}-\bar{g},\bar{G}) + Q^+(\bar{g},\bar{G}-\bar{g}) + \bar{G}\,L(\bar{G}-\bar{g}), \\ (\bar{G}-\bar{g})(t=0,.) = 0. \end{cases} \tag{5.14}$$

As before, we define the norm relative to ξ for all φ by

$$\|\varphi\|_t = \operatorname*{ess\,sup}_{s\in[0,t]\; x\in\mathbb{R}^N\; v\in\mathbb{R}^N} \frac{|\varphi(s,x,v)|}{\xi(s,x,v)}.$$

Using $\dot{\xi}^{\#} = 0$, the equation (5.14) gives

$$\Big(\frac{\bar{G}-\bar{g}}{\xi}\Big)^{\#}(t,x,v) \leq \int_0^t \frac{1}{\xi^{\#}}\big(Q^+(\bar{G}-\bar{g},\bar{G}) + Q^+(\bar{g},\bar{G}-\bar{g}) + \bar{G}\,L(\bar{G}-\bar{g})\big)^{\#}(s,x,v)\,ds. \tag{5.15}$$

Again, multiplying and dividing the collision terms by ξ and using (5.4) we deduce from (5.15) the following inequality:

$$\Big(\frac{\bar{G}-\bar{g}}{\xi}\Big)^{\#}(t,x,v) \leq 3\int_0^t \|\bar{G}\|_s\,\|\bar{G}-\bar{g}\|_s\,L(\xi)^{\#}\,ds$$

so that

$$\|\bar{G}-\bar{g}\|_T \leq 3\int_0^T \frac{C(t)}{(1+t)^N}\,\|\bar{G}-\bar{g}\|_t\,dt \qquad \text{for all } T>0.$$

Thanks to the Gronwall lemma we prove that $\bar{G} = \bar{g}$ is a distributional solution of the Boltzmann equation. $\square$

REFERENCES

[Ba,D,G] C. Bardos, P. Degond, and F. Golse, *A priori estimates and existence results for the Vlasov and Boltzmann equations*, in Proc. 1984 AMS-SIAM Summer Seminar, Santa Fe, NM.

[B,P,T] N. Bellomo, A. Pelczewski, and G. Toscani, *Mathematical Topics in Nonlinear Kinetic Theory*, World Scientific, Singapore.

[B,T] N. Bellomo and G. Toscani, *On the Cauchy problem for the nonlinear Boltzmann equation. Global existence, uniqueness and asymptotic stability*, J. Math. Phys., 26 (1985), pp. 334–338.

[C] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer-Verlag, Berlin, 1988.

[C,I,P] C. Cercignani, R. Illner, and M. Pulvirenti, *The mathematical theory of dilute gases*, Appl. Math. Sci., Springer-Verlag, Berlin, 1994.

[DP,L1] R. J. DiPerna and P-L. Lions, *On the Cauchy problem for Boltzmann equations: Global existence and weak stability*, Ann. Math., 130 (1989), pp. 321–366.

[DP,L2] 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.

[G,L,P,S] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solutions of a transport equation*, J. Functional Anal., 76 (1988), pp. 110–125.

[Gr] H. Grad, *Principles of the kinetic theory of gases*, Flügge's Handbuch der Physik, 12 (1958), pp. 205–294.

[K,S] S. Kaniel and M. Shinbrot, *The Boltzmann equation* I*: uniqueness and global existence*, Comm. Math. Phys., 58 (1978), pp. 65–84.

[H] K. Hamdache, *Existence in the large and asymptotic behaviour for the Boltzmann equation*, Japan J. Appl. Math., 2 (1985), pp. 1–15.

[I,S] R. Illner and M. Shinbrot, *Global existence for a rare gas in an infinite vacuum*, Comm. Math. Phys., 95 (1984), pp. 117–126.

[L1,L2] P.-L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications. Parts* I *and* II, J. Math. Kyoto Univ., 34 (1994), pp. 1–61.

[L3] P.-L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications. Part* III, J. Math. Kyoto Univ., 34 (1994), pp. 539–584.

[L 4] P.-L. Lions, *Conditions at infinity for Boltzmann's equation*, Comm. Partial Differential Equations, 19 (1994), pp. 335–367.

[L P] P.-L. Lions, B. Perthame, *Lemmes de moments, de moyenne et de dispersion*, C. R. Acad. Sci. Paris, 314 (1992), pp. 801–806.

[Pa] A. Palczewski, *Existence of global solutions to the Boltzmann equation in L^∞*, in Rarefied Gas Dynamics, V. Boffi, C. Cercignani, eds., Teubner, Stuttgart; *Existence in the large and asymptotic behaviour for the Boltzmann equation*, 1 (1986), pp. 144–149.

[P1] B. Perthame, *Time Decay, Propagation of Low Moments and Dispersive Effects for Kinetic Equations*, Comm. Partial Differential Equations, 21 (1996), pp. 659–686.

[P2] B. Perthame, *Global existence to the B.G.K. model of Boltzmann equation*, J. Differential Equations, 82 (1989), pp. 191–205.

[T] G. Toscani, *On the nonlinear Boltzmann equation in unbounded domains*, Arch. Rational Mech. Anal., 95 (1986), pp. 37–49.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1028–1047, September 1997

002

CLASSICAL SOLUTIONS OF MULTIDIMENSIONAL HELE–SHAW MODELS [*]

JOACHIM ESCHER[†] AND GIERI SIMONETT[‡]

Abstract. Existence and uniqueness of classical solutions for the multidimensional expanding Hele–Shaw problem are proved.

Key words. classical solutions, Hele–Shaw model, moving boundary problem, maximal regularity

AMS subject classifications. 35R35, 35K55, 35S30, 76D99

PII. S0036141095291919

1. The problem.

We are concerned with a class of moving boundary problems for bounded domains in $\mathbb{R}^n$, which comprise in particular the so-called single phase Hele–Shaw problem. In order to describe precisely the involved geometry, let Ω be a bounded domain in $\mathbb{R}^n$ and assume that its boundary $\partial\Omega$ is of class C^∞. Moreover, assume that $\partial\Omega$ consists of two disjoint nonempty components J and Γ. Later on, we will model over the exterior component Γ a moving interface, whereas the interior component J describes a fixed portion of the boundary. Let ν denote the outer unit normal field over Γ and fix $\alpha \in (0,1)$. Given $a > 0$, set

$$\mathcal{U} := \{\rho \in C^{2+\alpha}(\Gamma)\,;\, \|\rho\|_{C^1(\Gamma)} < a\}.$$

For each $\rho \in \mathcal{U}$ define the map

$$\theta_\rho := id_\Gamma + \rho\nu$$

and let $\Gamma_\rho := \operatorname{im}(\theta_\rho)$ denote its image. Obviously, θ_ρ is a $C^{2+\alpha}$ diffeomorphism mapping Γ onto Γ_ρ, provided $a > 0$ is chosen sufficiently small. In addition, we assume that $a > 0$ is small enough such that Γ_ρ and J are disjoint for each $\rho \in \mathcal{U}$. Let Ω_ρ denote the domain in $\mathbb{R}^n$ being diffeomorphic to Ω and whose boundary is given by J and Γ_ρ. To describe the evolution of the hypersurface Γ_ρ, fix $T > 0$ and set $I := [0, T]$. Then each map $\rho : I \to \mathcal{U}$ defines a collection of domains $\Omega_{\rho(t)},\ t \in I$. For later purposes it is convenient to introduce the following generalized parabolic cylinder:

$$\Omega_{\rho,T} := \big\{(x,t) \in \mathbb{R}^n \times [0,T]\,;\, x \in \Omega_{\rho(t)}\big\} = \bigcup_{t\in I}\big(\Omega_{\rho(t)} \times \{t\}\big)$$

and, correspondingly,

$$\Gamma_{\rho,T} := \big\{(x,t) \in \mathbb{R}^n \times [0,T]\,;\, x \in \Gamma_{\rho(t)}\big\} = \bigcup_{t\in I}\big(\Gamma_{\rho(t)} \times \{t\}\big).$$

[*] Received by the editors September 15, 1995; accepted for publication (in revised form) April 17, 1996.

http://www.siam.org/journals/sima/28-5/29191.html

[†] Mathematical Institute, University of Basel, CH-4051 Basel, Switzerland (escher@math.unibas.ch).

[‡] Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (simonett@math.vanderbilt.edu). The research of this author was supported by Schweizerischer Nationalfonds.

Observe that $\Omega_{0,T}$ is just the standard parabolic cylinder $\Omega \times [0,T]$. Similarly, $\Gamma_{0,T} = \Gamma \times [0,T]$. For the sake of completeness, we write $J_T := J \times [0,T]$.

Now let $\rho_0 \in \mathcal{U}$ be given. Moreover, pick $b \in C(J)$ and $\delta \in \{0,1\}$. Then we consider the *moving boundary problem* of determining a pair (u,ρ) satisfying the following set of equations:

$$(1.1)_{\rho_0} \qquad \begin{aligned} \Delta u &= 0 \quad &&\text{in} \quad \Omega_{\rho,T}, \\ u &= 0 \quad &&\text{on} \quad \Gamma_{\rho,T}, \\ (1-\delta)u + \delta(\nabla u|\nu_J) &= b \quad &&\text{on} \quad J_T, \\ \partial_t N_\rho - (\nabla u|\nabla N_\rho) &= 0 \quad &&\text{on} \quad \Gamma_{\rho,T}, \\ \rho(0,\cdot) &= \rho_0 \quad &&\text{on} \quad \Gamma. \end{aligned}$$

Here, Δ and ∇ stand for the Laplacian and the gradient, respectively, in the Euclidean metric. The outer unit normal field over J is denoted by ν_J. The parameter δ is introduced to label the boundary condition on the fixed boundary J (where $\delta = 0$ corresponds to a Dirichlet boundary condition and $\delta = 1$ corresponds to a Neumann condition). Moreover, N_ρ is a defining function for Γ_ρ, i.e., $\Gamma_\rho = N_\rho^{-1}(0)$, $\rho \in \mathcal{U}$. A precise definition of N_ρ is given in section 2.

The set of equations in (1.1) express that the free boundary moves with normal velocity given by the normal derivative of a harmonic function which vanishes on the boundary. More precisely, the motion of the free boundary is governed by $V = -\frac{\partial u}{\partial \nu}$, where the function u satisfies the first three equations in (1.1). Here, V is the normal velocity taken to be positive for expanding hypersurfaces and ν is the outer unit normal field on the moving boundary.

Assume now that $n = 2$, $\delta = 1$, and $b > 0$. Then problem $(1.1)_{\rho_0}$ represents the classical formulation of the expanding two-dimensional Hele–Shaw flow; see Crank [5], Elliott and Ockendon [10], Elliott and Janovsky [9], DiBenedetto and Friedman [7], and Richardson [21]. In this model, u has the meaning of the pressure in an incompressible viscous fluid blob Ω_ρ. Since b is positive, further fluid is injected through the fixed boundary J at the rate b. Hence, the blob is advancing in time, modelled by the moving boundary Γ_ρ. Some authors (see Fasano and Primicerio [15] or Steinbach and Weinelt [22]) consider the above model in the case of prescribed pressure on the fixed boundary, i.e., with the inhomogeneous Dirichlet boundary condition $u = b$ on J. This boundary condition corresponds to the case $\delta = 0$ in $(1.1)_{\rho_0}$. In our model, we cover both cases and we prove the existence of a unique *classical* solution (u,ρ) for the general problem $(1.1)_{\rho_0}$; see the main result below. As pointed out in [5], [8], [9], [10], [16], and [22], there are further applications of $(1.1)_{\rho_0}$ to different multidimensional moving boundary problems. We mention the electrochemical machining problem, the one-phase Stefan problem with zero specific heat, the flow of viscous fluid through porous media, and the injection moulding process. These models make sense in higher space dimensions and under general boundary conditions on the fixed boundary J.

To clearly state our result, we need some definitions. Given an open subset U of $\mathbb{R}^m$, let $h^s(U)$ denote the little Hölder space of order $s > 0$, a closed subspace of the usual Hölder space $BUC^s(U)$; see section 2 for a precise definition. Throughout this paper we fix $\alpha \in (0,1)$ and we define

$$\mathcal{V} := h^{2+\alpha}(\Gamma) \cap \mathcal{U}.$$

Moreover, we need the anisotropic function spaces $Ch^{0,s}(\Omega_{\rho,T})$ consisting of all $u : \Omega_{\rho,T} \to \mathbb{R}$ such that, given $(x,t) \in \Omega_{\rho,T}$, the function $u(\cdot,t)$ belongs to $h^s(\Omega_{\rho(t)})$ and the function $u(x,\cdot)$ belongs to $C([0,T])$. A pair (u,ρ) is called a *classical Hölder solution* of (1.1) if

$$(u,\rho) \in Ch^{0,2+\alpha}(\Omega_{\rho,T}) \times \big(C([0,T),\mathcal{V}) \cap C^1([0,T),h^{1+\alpha}(\Gamma))\big)$$

and if (u,ρ) satisfies the equations in (1.1) pointwise. Our main result now reads as follows.

THEOREM 1.1. *Assume that $b \in h^{2+\alpha-\delta}(J)$ is nonnegative and not identically equal to zero. Then, given any initial value $\rho_0 \in \mathcal{V}$, there exist $T > 0$ and a unique classical solution (u,ρ) of $(1.1)_{\rho_0}$ on $[0,T]$. Moreover, the moving boundary $\rho : (0,T) \to \mathcal{V}$ is analytic in the time variable.*

It should be emphasized that Theorem 1.1 guarantees a unique classical solution to problem (1.1) for each $C^{2+\alpha}$ initial hypersurface Γ_{ρ_0} which is close to Γ in the sense that ρ_0 belongs to $\mathcal{V}$.

In Elliott [8] and Elliott and Janovsky [9], a variational inequality approach for problem $(1.1)_{\rho_0}$ is developed, and the existence and uniqueness of global weak solutions are proved. However, as stated in the Conclusion of [9] (see p. 106), the existence of classical solutions left an open problem.

Our approach to problem $(1.1)_{\rho_0}$ proposed in this paper is of a different nature. Indeed, transforming the original problem on a fixed domain, we are looking for classical solutions from the very beginning. After a natural reduction of the transformed equations, we are led to an evolution equation for the moving boundary involving a nonlinear and nonlocal pseudodifferential operator of first order. The main result for this pseudodifferential operator can be summarized by the fact that it depends smoothly on the unknown and that the corresponding linearized operator is a nicely behaving operator; i.e., it generates a strongly continuous analytic semigroup on an appropriate subspace of Hölder continuous functions, provided $b \geq 0$ and $b \neq 0$. This generation property of the linearization makes it possible to use the general results of the theory of maximal regularity, due to Da Prato and Grisvard [6], and to construct a unique classical solution of the nonlinear problem. The same technique has been applied to moving boundary problems arising in gravity flows of incompressible fluids through porous media; see [12] and [13].

There is a one-dimensional version of problem $(1.1)_{\rho_0}$; see the work of Fasano and Primicerio [14], [15]. Since the geometry of one-dimensional moving boundary problems is considerably easier to handle, classical solutions are well known to exist in this case.

For two-dimensional simply connected domains and for initial data belonging to an appropriate Gevrey class, Reissig [20] recently proved the existence of analytic solutions to a Hele–Shaw model with a point source.

Let $\rho_0 \in \mathcal{V}$ be given and assume that $b \in h^{2+\alpha-\delta}(J) \setminus \{0\}$ is nonnegative. Moreover, let (u,ρ) denote the classical solution of $(1.1)_{\rho_0}$ constructed in Theorem 1. Then, given $t \in [0,T]$, the pressure $u(\cdot,t) \in h^{2+\alpha}(\Omega_{\rho(t)})$ is the unique solution in $h^{2+\alpha}(\Omega_{\rho(t)})$ of the following elliptic boundary value problem:

$$\Delta u = 0 \quad \text{in} \quad \Omega_{\rho(t)}, \quad u = 0 \quad \text{on} \quad \Gamma_{\rho(t)}, \quad (1-\delta)u + \delta(\nabla u | \nu_J) = b \quad \text{on} \quad J.$$

Hence the strong maximum principle implies that the pressure $u(\cdot,t)$ is strictly positive in $\Omega_{\rho(t)}$. This property is crucial for our approach; see step (b) in the proof of Theorem 4.2.

From a mathematical and a physical point of view, problem $(1.1)_{\rho_0}$ also makes sense for negative b. However, in this so-called ill-posed case, the problem has a completely different feature, as pointed out by Elliott and Ockendon [10] based on numerical investigations, by DiBenedetto and Friedman [7] proving so-called fingering, and by Fasano and Primicerio [15] establishing blow-up and nonexistence results for one-dimensional problems. Our results are also optimal in this sense, since we guarantee the existence of classical solutions in the well-posed case $b \geq 0$, $b \neq 0$, and we *prove* that the linearized reduced problem for the moving boundary is ill-posed in the sense of Hadamard for $b \leq 0$, $b \neq 0$; see Remark 5.3.

2. The transformed problem.

In this section we transform the original problem into a problem on a fixed domain, and we introduce a nonlinear, nonlocal pseudo-differential operator Φ of an appropriate reduced problem for the moving boundary Γ_ρ. In addition, we provide a useful representation of the Fréchet derivative of Φ.

Let us first introduce some function spaces which we will need in what follows. Assume that U is an open subset of $\mathbb{R}^m$. Given $k \in \mathbb{N} \cup \{\infty\}$, let $C^k(U)$ denote the space of all $f : U \to \mathbb{R}$ having continuous derivatives up to order k. The closed subspace of $C^k(U)$ consisting of all maps from U into $\mathbb{R}$ which have bounded and uniformly continuous derivatives up to order k is denoted by $BUC^k(U)$. Given $\alpha \in (0,1)$, the space $BUC^{k+\alpha}(U)$ stands for all $f \in BUC^k(U)$ having uniformly α-Hölder continuous derivatives of order k. In addition, $C^\omega(U)$ denotes the subspace of all real analytic functions on U.

Furthermore, we write $\mathcal{S}(\mathbb{R}^m)$ for the Schwartz space, i.e., the Fréchet space of all rapidly decreasing smooth functions on $\mathbb{R}^m$.

Next let r_U denote the restriction operator with respect to U, i.e., $r_U u := u|U$ for $u \in BUC(U)$. Then the *little Hölder spaces* $h^s(U)$, $s \geq 0$, are defined as

$$h^s(U) := \text{closure of } \ r_U\big(\mathcal{S}(\mathbb{R}^m)\big) \ \text{ in } \ BUC^s(U).$$

Finally, assume that M is an m-dimensional (sufficiently) smooth submanifold of $\mathbb{R}^n$. Then the spaces $BUC^s(M)$ and $h^s(M)$, $s \geq 0$, are defined as usual by means of a smooth atlas for M; see [24].

It is useful to write Γ_ρ as a 0-level set of an appropriate function. For this, pick $a_0 \in (0, \operatorname{dist}(\Gamma, J))$ and let

$$\mathcal{N} : \Gamma \times (-a_0, a_0) \to \mathbb{R}^n, \qquad \mathcal{N}(x, \lambda) := x + \lambda\nu(x).$$

If $a_0 > 0$ is small enough, we have that

$$\mathcal{N} \in \operatorname{Diff}^\infty(\Gamma \times (-a_0, a_0), \mathcal{R}),$$

where $\mathcal{R} := \operatorname{im}(\mathcal{N})$. It is convenient to decompose the inverse of $\mathcal{N}$ into $\mathcal{N}^{-1} = (X, \Lambda)$, where

$$X \in BUC^\infty(\mathcal{R}, \Gamma) \qquad \text{and} \qquad \Lambda \in BUC^\infty(\mathcal{R}, (-a_0, a_0)).$$

Note that $X(y)$ is the nearest point on Γ to y and that $\Lambda(y)$ is the signed distance from y to Γ (that is, to $X(y)$). The neighborhood $\mathcal{R}$ consists of those points with distance less than a_0 to Γ. Given $\rho \in \mathcal{V}$, now define

$$N_\rho : \mathcal{R} \to \mathbb{R}, \qquad N_\rho(y) := \Lambda(y) - \rho(X(y)).$$

Then it is not difficult to verify that $\Gamma_\rho = N_\rho^{-1}(0)$. Therefore, the gradient ∇N_ρ is perpendicular to Γ_ρ, and ∇N_ρ points outward since $N_\rho(y) < 0$ if $y \in \Omega_\rho$. So it follows that the outer unit normal field ν on Γ_ρ is given by $\nu = \frac{\nabla N_\rho}{|\nabla N_\rho|}$. Let $\rho \in C^1([0,T], h^{1+\alpha}(\Gamma))$ be given and set

$$N_\rho(y,t) := \Lambda(y) - \rho(X(y),t), \qquad y \in \mathcal{R}, \quad t \in [0,T].$$

Then

$$V(y,t) := -\frac{\partial_t N_\rho(y,t)}{|\nabla N_\rho(y,t)|} = \frac{\partial_t \rho(X(y),t)}{|\nabla N_\rho(y,t)|}, \qquad y \in \Gamma_{\rho(t)}, \quad t \in [0,T],$$

is the normal velocity of the moving hypersurfaces $\Gamma_{\rho(t)}$ in the direction of the outer normal field. Hence the fourth equation in (1.1) can be rewritten as $-\frac{\partial_t N_\rho}{|\nabla N_\rho|} = -(\nabla u|\nu)$, which shows that the motion of the hypersurfaces $\Gamma_{\rho(t)}$ is governed by $V = -\frac{\partial u}{\partial \nu}$.

Next we introduce an appropriate extension of θ_ρ to $\mathbb{R}^n$. For this we assume that $a \in (0, a_0/4)$, and we fix a $\varphi \in C^\infty(\mathbb{R}, [0,1])$ such that

$$\varphi(\lambda) = \begin{cases} 1 & \text{if} \quad |\lambda| \le a, \\ 0 & \text{if} \quad |\lambda| \ge 3a \end{cases}$$

and such that $\sup |\partial \varphi(\lambda)| < 1/a$. Then we define for each $\rho \in \mathcal{V}$ the map

$$\Theta_\rho(y) := \begin{cases} \mathcal{N}\big(X(y), \Lambda(y) + \varphi(\Lambda(y))\rho(X(y))\big) & \text{if} \quad y \in \mathcal{R}, \\ y & \text{if} \quad y \notin \mathcal{R}. \end{cases}$$

Note that $[\lambda \mapsto \lambda + \varphi(\lambda)\rho]$ is strictly increasing since $|\partial\varphi(\lambda)\rho| < 1$. Then it is not difficult to verify that

$$\Theta_\rho \in \mathrm{Diff}^{\,2+\alpha}(\mathbb{R}^n, \mathbb{R}^n) \cap \mathrm{Diff}^{\,2+\alpha}(\Omega, \Omega_\rho) \quad \text{and} \quad \Theta_\rho | \Gamma = \theta_\rho.$$

Moreover, we observe that there exists an open neighborhood U of J such that

$$\Theta_\rho | U = id_U. \tag{2.1}$$

It should be mentioned that the above diffeomorphism was first introduced by Hanzawa [18] to transform multidimensional Stefan problems to fixed domains. In the following we use the same symbol θ_ρ for both diffeomorphisms θ_ρ and Θ_ρ. The pull-back operator induced by θ_ρ is given as

$$\theta^* u := \theta^*_\rho u := u \circ \theta_\rho \quad \text{for} \quad u \in BUC(\Omega_\rho).$$

Similarly, the corresponding push-forward operator is defined as

$$\theta_* v := \theta^\rho_* v := v \circ \theta_\rho^{-1} \quad \text{for} \quad v \in BUC(\Omega).$$

Lemma 2.1. *Given $\rho \in \mathcal{V}$ and $k \in \{1, 2\}$, we have*

$$\theta^*_\rho \in \mathrm{Isom}(h^{k+\alpha}(\Omega_\rho), h^{k+\alpha}(\Omega)) \cap \mathrm{Isom}(h^{k+\alpha}(\Gamma_\rho), h^{k+\alpha}(\Gamma))$$

with $[\theta_\rho^*]^{-1} = \theta_*^\rho$.

Proof. Let $\rho \in \mathcal{V}$ and $k \in \{1, 2\}$ be given. It follows from the mean value theorem that

$$\theta_\rho^* \in \mathrm{Isom}(BUC^{k+\alpha}(\Omega_\rho), BUC^{k+\alpha}(\Omega)).$$

Hence, to prove the first assertion, it suffices to show that $\theta_\rho^* u$ belongs to the space $h^{k+\alpha}(\Omega)$, whenever u belongs to $h^{k+\alpha}(\Omega_\rho)$. But this is an easy consequence of the following known characterization of little Hölder spaces: a function $u \in BUC^{k+\alpha}(\Omega)$ belongs to $h^{k+\alpha}(\Omega)$ iff

$$\lim_{\tau\to 0^+} \sup_{0<|x-y|\le\tau} \frac{|\partial^\beta u(x) - \partial^\beta u(y)|}{\tau^\alpha} = 0, \qquad \beta \in \mathbb{N}^n,\ |\beta| = k.$$

This can be seen by means of local coordinate charts along the lines of Lemma 2.7 and Remark 2.8 in [19]; see also [3]. The second assertion follows analogously. ∎

Given $\rho \in \mathcal{V}$, we now introduce the following transformed differential operators, acting linearly on $BUC^2(\Omega)$:

$$\begin{aligned} &A(\rho)v := -\theta_\rho^*\big(\Delta(\theta_*^\rho v)\big), \qquad B(\rho)v := \gamma\theta_\rho^*(\nabla(\theta_*^\rho v)|\nabla N_\rho), \\ &Cv := (1-\delta)\gamma_J v + \delta(\gamma_J \nabla v|\nu_J), \end{aligned}$$

where γ and γ_J denote the trace operators with respect to Γ and J, respectively. Assume now that (u, ρ) is a classical Hölder solution of $(1.1)_{\rho_0}$. Then it is not difficult to see that $v := [t \mapsto \theta_{\rho(t)}^* u(t, \cdot)]$ belongs to $C([0, T], h^{2+\alpha}(\Omega))$ and that the pair (v, ρ) satisfies the following equations:

$$(2.2)_{\rho_0} \qquad \begin{aligned} A(\rho)v &= 0 \quad \text{in} \quad \Omega_{0,T}, \\ v &= 0 \quad \text{on} \quad \Gamma_{0,T}, \\ Cv &= b \quad \text{on} \quad J_T, \\ \partial_t\rho + B(\rho)v &= 0 \quad \text{on} \quad \Gamma_{0,T}, \\ \rho(0, \cdot) &= \rho_0 \quad \text{on} \quad \Gamma. \end{aligned}$$

A pair (v, ρ) is called a *classical Hölder solution* of $(2.2)_{\rho_0}$ if

$$\begin{aligned} &v \in C([0, T], h^{2+\alpha}(\Omega)), \\ &\rho \in C([0, T], \mathcal{V}) \cap C^1([0, T], h^{1+\alpha}(\Gamma)) \end{aligned}$$

and if (v, ρ) satisfies the equations in $(2.2)_{\rho_0}$ pointwise. The following lemma is an obvious consequence of Lemma 2.1 and (2.1).

LEMMA 2.2. *Let* $\rho_0 \in \mathcal{V}$ *be given.*

(a) *If* (u, ρ) *is a classical Hölder solution of* $(1.1)_{\rho_0}$, *then* $(\theta_\rho^* u, \rho)$ *is a classical Hölder solution of* $(2.2)_{\rho_0}$.

(b) *If* (v, ρ) *is a classical Hölder solution of* $(2.2)_{\rho_0}$, *then* $(\theta_*^\rho v, \rho)$ *is a classical Hölder solution of* $(1.1)_{\rho_0}$.

In the next two lemmas we collect some results for elliptic boundary value problems in little Hölder spaces. We shall use these results in sections 3 and 4.

Lemma 2.3.

$$(A, B) \in C^{\omega}(\mathcal{V}, \mathcal{L}(h^{2+\alpha}(\Omega), h^{\alpha}(\Omega) \times h^{1+\alpha}(\Gamma))).$$

Proof. Let η denote the standard Euclidean metric on $\mathbb{R}^m$ and let $\theta^*\eta$ be the Riemannian metric on $\overline{\Omega}$ induced by the diffeomorphism θ_ρ, i.e.,

$$\theta_\rho^*\eta|_x(\xi, \zeta) := \eta|_{\theta_\rho(x)}(T_x\theta_\rho\xi, T_x\theta_\rho\zeta)$$

for $x \in \overline{\Omega}$ and ξ, $\zeta \in T_x(\overline{\Omega})$. Then $A(\rho)$ and $B(\rho)$ are just the Laplace–Beltrami operator and the outer normal derivative of $(\Omega, \theta_\rho^*\eta)$. Since the metric $\theta_\rho^*\eta$ depends analytically on $\rho \in \mathcal{V}$, the assertion follows easily. □

Lemma 2.4. *Let $\rho \in \mathcal{V}$ be given. Then for each*

$$(f, g, h) \in h^{\alpha}(\Omega) \times h^{2+\alpha}(\Gamma) \times h^{2+\alpha-\delta}(J)$$

there exists a unique classical solution $v := V(\rho)(f, g, h)$ in $h^{2+\alpha}(\Omega)$ of

$$A(\rho)v = f \quad \textit{in} \quad \Omega, \qquad v = g \quad \textit{on} \quad \Gamma, \qquad Cv = h \quad \textit{on} \quad J.$$

Moreover, there exists a positive constant $C := C(\rho)$ such that

$$\|V(\rho)(f, g, h)\|_{2+\alpha,\Omega} \le C\big(\|f\|_{\alpha,\Omega} + \|g\|_{2+\alpha,\Gamma} + \|h\|_{2+\alpha-\delta,J}\big).$$

Proof. (a) It follows from the proof of Lemma 2.3 and by construction that A is a uniformly elliptic operator having α-Hölder continuous coefficients and that C is a normal boundary operator with regular coefficients too. Hence we conclude from Theorem 7.3 and Remark 2 on p. 669 in [1] that, given any compact subset K of $\mathcal{V}$, there exists a positive constant $C := C(K)$ such that

$$\|v\|_{2+\alpha,\Omega} \le C\big(\|A(\rho)v\|_{\alpha,\Omega} + \|\gamma v\|_{2+\alpha,\Gamma} + \|Cv\|_{2+\alpha-\delta,J}\big)$$

for all $v \in h^{2+\alpha}(\Omega)$ and all $\rho \in K$.

(b) Observe that $\big(A(0), \gamma, C\big)$ is a regular elliptic boundary value problem with constant coefficients on a smooth domain. Hence it follows from formula (3) on p. 236 in [24] that

$$\big(A(0), \gamma, C\big) \in \mathrm{Isom}(h^{2+\alpha}(\Omega), h^{\alpha}(\Omega) \times h^{2+\alpha}(\Gamma) \times h^{2+\alpha-\delta}(J)).$$

Now let $\rho \in \mathcal{V}$ be given and set $K := \{t\rho\,;\, t \in [0, 1]\}$. Then K is a compact subset of $\mathcal{V}$, and therefore it follows from (a) and the continuity method (see Theorem 5.2 in [17]) that

$$\big(A(\rho), \gamma, C\big) \in \mathrm{Isom}(h^{2+\alpha}(\Omega), h^{\alpha}(\Omega) \times h^{2+\alpha}(\Gamma) \times h^{2+\alpha-\delta}(J)).$$

This completes our argumentation. □

Let us now introduce the natural decomposition $V = S \oplus T \oplus R$ of the above solution operator by setting

$$\begin{aligned} S(\rho) &:= V(\rho)(\cdot, 0, 0) \in \mathcal{L}(h^{\alpha}(\Omega), h^{2+\alpha}(\Omega)), \\ T(\rho) &:= V(\rho)(0, \cdot, 0) \in \mathcal{L}(h^{2+\alpha}(\Gamma), h^{2+\alpha}(\Omega)), \\ R(\rho) &:= V(\rho)(0, 0, \cdot) \in \mathcal{L}(h^{2+\alpha-\delta}(J), h^{2+\alpha}(\Omega)). \end{aligned}$$

Given $v \in BUC^1(\Omega)$, let $\partial_\nu v$ denote the directional derivative with respect to the outer unit normal on Γ, i.e., $\partial_\nu v := \gamma(\nabla v|\nu)$. Using this notation it follows from the strong maximum principle that

$$\partial_\nu(R(\rho)b) < 0, \tag{2.3}$$

provided $b \in h^{2+\alpha-\delta}(J) \setminus \{0\}$ with $b \geq 0$.

Throughout the remainder of this paper we fix

$$b \in h^{2+\alpha-\delta}(J) \setminus \{0\} \quad \text{with} \quad b \geq 0 \tag{2.4}$$

and we set

$$\Phi(\rho) := B(\rho)R(\rho)b \quad \text{for} \quad \rho \in \mathcal{V}.$$

It follows from Lemma 2.3 and the definition of R that Φ maps $\mathcal{V}$ into $h^{1+\alpha}(\Gamma)$. Given $\rho_0 \in \mathcal{V}$, we now consider the nonlinear evolution equation in $h^{1+\alpha}(\Gamma)$ for the operator Φ:

$$\partial_t \rho + \Phi(\rho) = 0, \qquad \rho(0) = \rho_0. \tag{2.5}$$

A function $\rho : I = [0,T] \to h^{1+\alpha}(\Gamma)$ is called a *classical Hölder solution* of (2.5) if

$$\rho \in C(I, \mathcal{V}) \cap C^1(I, h^{1+\alpha}(\Gamma))$$

and if ρ satisfies (2.5) pointwise on I. Using this notation it is now easy to state the following *reduction* of the transformed problem (2.2).

LEMMA 2.5. *Let $\rho_0 \in \mathcal{V}$ be given.*

(a) *If ρ is a classical Hölder solution of* (2.5)*, then the pair $(R(\rho)b, \rho)$ is a classical Hölder solution of* (2.2)*.*

(b) *Suppose that (v, ρ) is a classical Hölder solution of* (2.2)*. Then ρ is a classical Hölder solution of* (2.5)*.*

Proof. This follows immediately from the definition of $R(\rho)$. □

In order to treat the nonlinear evolution equation (2.5), we first show that $\Phi(\rho)$ depends smoothly on $\rho \in \mathcal{V}$ and we provide an appropriate representation of the Fréchet derivative $\partial\Phi(\rho)$ of Φ at $\rho \in \mathcal{V}$. For this we introduce for each $\rho \in \mathcal{V}$ the following linear operators:

$$\begin{aligned} K &:= K(\rho) := -\partial A(\rho)[\cdot, R(\rho)b] \in \mathcal{L}(h^{2+\alpha}(\Gamma), h^\alpha(\Omega)), \\ M &:= M(\rho) := \partial B(\rho)[\cdot, R(\rho)b] \in \mathcal{L}(h^{2+\alpha}(\Gamma), h^{1+\alpha}(\Gamma)). \end{aligned}$$

Here, the notation $\partial A(\rho)[h,v]$ stands for

$$\partial A(\rho)[h,v] = \frac{\partial}{\partial\varepsilon}\bigg|_{\varepsilon=0} A(\rho + \varepsilon h)v, \qquad h \in h^{2+\alpha}(\Gamma),\ v \in h^{2+\alpha}(\Omega).$$

LEMMA 2.6. $\Phi \in C^\omega(\mathcal{V}, h^{1+\alpha}(\Gamma))$ *with*

$$\partial\Phi(\rho) = B(\rho)S(\rho)K(\rho) + M(\rho)$$

for each $\rho \in \mathcal{V}$.

Proof. (a) Due to Lemma 2.3, it suffices to show that

$$[\rho \mapsto R(\rho)b] \in C^\omega(\mathcal{V}, h^{2+\alpha}(\Omega)) \quad \text{with} \quad \partial(R(\rho)b) = S(\rho)K(\rho).$$

(b) Recall that $\mathcal{V}$ is an open subset of $h^{2+\alpha}$. Let γ denote the trace operator with respect to Γ and let

$$F(\rho, v) := (A(\rho)v, \gamma v, Cv - b), \qquad (\rho, v) \in \mathcal{V} \times h^{2+\alpha}(\Omega).$$

Then it follows from Lemma 2.3 that

$$F \in C^\omega\big(\mathcal{V} \times h^{2+\alpha}(\Omega), h^\alpha(\Omega) \times h^{2+\alpha}(\Gamma) \times h^{2+\alpha-\delta}(J)\big).$$

Moreover, given $(\rho, v) \in \mathcal{V} \times h^{2+\alpha}(\Omega)$, we have that

$$\partial_2 F(\rho, v)w = (A(\rho)w, \gamma w, Cw) \quad \text{and} \quad \partial_1 F(\rho, v)h = (\partial A(\rho)[h, v], 0, 0)$$

for $w \in h^{2+\alpha}(\Omega)$ and $h \in h^{2+\alpha}(\Gamma)$. Now the assertion follows from Lemma 2.4 and the implicit function theorem. □

The next two sections are devoted to the study of the linearization $\partial\Phi(\rho)$ of Φ. We will see that it is a nicely behaving operator; i.e., we will prove that $-\partial\Phi(\rho)$ generates a strongly continuous analytic semigroup on $h^{1+\alpha}(\Gamma)$.

3. Localizations. Given $\kappa \in (0, a]$, let $\mathcal{R}_\kappa := \mathcal{N}(\Gamma \times (-\kappa, 0])$. Then there exists $m := m_\kappa \in \mathbb{N}$ and an atlas $\{(U_l, \varphi_l) \,;\, 1 \le l \le m\}$ of $\mathcal{R}_\kappa$ such that $\operatorname{diam}(U_l) < 2\kappa$ for all $l \in \{1, \dots, m\}$. Let

$$s_l \in C^\infty((-\delta, \delta)^{n-1}, U_l), \quad l \in \{1, \dots, m\},$$

be a parameterization of $U_l \cap \Gamma$. Furthermore, let $P := (-\delta, \delta)^{n-1}$ and $Q := P \times [0, \delta)$ and define

$$\mu_l : Q \to U_l, \qquad (\omega, r) \mapsto s_l(\omega) - r\nu(s_l(\omega)).$$

Without loss of generality, we may assume that $\delta = \kappa$ and that $\mu_l := \varphi_l^{-1}$ for $1 \le l \le m$. The additional parameter κ is introduced to control the size of the chart domain U_l. This fact will be used in section 5 to prove a perturbation result; cf. Lemma 5.1. Finally, to further economize our notation, we set $\mu := \mu_l$, $U := U_l$ and we let

$$\mu^* u := u \circ \mu, \quad u \in C(U_l) \quad \text{and} \quad \mu_* v := v \circ \mu^{-1}, \quad v \in C(Q)$$

denote the pull-back and push-forward operators, respectively, induced by μ. Given $l \in \{1, \dots, m\}$, we define *local representations* $\mathcal{A} := \mathcal{A}_l$ and $\mathcal{B} := \mathcal{B}_l$ of A and B with respect to (Q, μ_l) by setting

$$\mathcal{A}(\mu^*\rho)\mu^* = \mu^* A(\rho) \quad \text{and} \quad \mathcal{B}(\mu^*\rho)\mu^* = \mu^* B(\rho), \quad \rho \in \mathcal{V},$$

respectively. To determine the coefficients of $\mathcal{A}$ and $\mathcal{B}$, let

$$\hat\rho := \hat\rho_l := \mu_l^* \rho, \quad \rho \in \mathcal{V}$$

and put $d(\omega, r) := \hat\rho(\omega) - r$ for $(\omega, r) \in Q$. In addition, we use the notation

$$\partial_j := \partial_{\omega_j}, \quad 1 \le j \le n-1, \qquad \partial_n := \partial_r.$$

Given $1 \leq j,\, k \leq n-1$, define

$$w_{jk} := (\partial_j s|\partial_k s) + d\big((\partial_j \mu^* \nu|\partial_k s) + (\partial_k \mu^* \nu|\partial_j s)\big) + d^2(\partial_j \mu^* \nu|\partial_k \mu^* \nu).$$

Clearly, $[w_{jk}]$ is symmetric. In addition, observe that $[(\partial_j s|\partial_k s)]$ is uniformly positive definite on P and that $\sup|d(\omega,r)| \leq 2a$. Hence we may assume also that $[w_{jk}]$ is uniformly positive definite on Q, provided $a>0$ is small enough. Let w denote the inverse of $[w_{jk}]$ and let w^{jk} be the components of w. Finally, set

$$D(\omega,r) := \begin{pmatrix} \nabla\hat{\rho}\otimes\nabla\hat{\rho} & \nabla\hat{\rho} \\ (\nabla\hat{\rho})^T & 1 \end{pmatrix}, \quad (\omega,r)\in Q,$$

and let

$$g_{jk} := g^l_{jk}(\rho) := (\partial_j \mu^* \theta_\rho|\partial_k \mu^* \theta_\rho), \quad 1 \leq j,\, k \leq n,$$

denote the components of the metric tensor with respect to (Q,μ). Note that

$$\mu^*\theta_\rho(\omega,r) = \theta_\rho(\mu(\omega,r)) = s(\omega) + d(\omega,r)\nu(s(\omega))$$

since $\varphi \equiv 1$ on $\mu(Q)$. In addition, observe that $d(w,r) = \hat{\rho}(\omega) - r$ is the function $-N_\rho$ in local coordinates. Using the orthogonality relations $(\partial_j s|\nu) = 0$ and $(\partial_j \nu|\nu) = 0$, direct calculations yield the formulas

$$[g_{jk}] = \begin{pmatrix} w^{-1} & 0 \\ 0 & 0 \end{pmatrix} + D \tag{3.1}$$

and

$$[g^{jk}] = \begin{pmatrix} w & -w\nabla\hat{\rho} \\ -(w\nabla\hat{\rho})^T & 1 + (w\nabla\hat{\rho}|\nabla\hat{\rho}) \end{pmatrix}, \tag{3.2}$$

where $[g^{jk}]$ is the inverse of $[g_{jk}]$. From (3.1), (3.2), and the well-known formula (which essentially is Cramer's rule)

$$g_{nn} = \det{[g_{jk}]_{1\leq j,k\leq n}} \cdot \det{[g^{jk}]_{1\leq j,k\leq n-1}},$$

one then deduces that

$$G := \sqrt{\det{[g_{jk}]_{1\leq j,k\leq n}}} = \sqrt{\det w^{-1}}. \tag{3.3}$$

Finally, let W denote the uniformly elliptic second-order differential operator acting on $C^2(P)$ which is induced by w, i.e.,

$$W\sigma := -\sum_{j,k=1}^{n-1} w^{jk}\partial_j\partial_k\sigma, \qquad \sigma\in C^2(P).$$

In the next lemma, we use the following notation: given $\widetilde{a} \in C^\infty(Q\times\mathbb{R}\times\mathbb{R}^{n-1},\mathbb{R})$ and $\sigma\in C^1(P)$, let $a(\sigma,\nabla\sigma)$ denote the Nemitskii operator induced by $\widetilde{a}$, i.e.,

$$a(\sigma,\nabla\sigma)(\omega,r) := \widetilde{a}((\omega,r),\sigma(\omega),\nabla\sigma(\omega)), \qquad (\omega,r)\in Q.$$

Lemma 3.1. *There exist*

$$\widetilde{a}_{jk},\ \widetilde{a}_j,\ \widetilde{b}_j \in C^\infty(Q\times(-a,a)\times\mathbb{R}^{n-1},\mathbb{R}),\qquad 1\le j,\ k\le n,$$

such that

$$\begin{aligned}&[\widetilde{a}_{jk}] \quad \textit{is symmetric and uniformly positive definite,}\\ &\widetilde{b}_n \quad \textit{is uniformly positive}\end{aligned}\tag{3.4}$$

on compact subsets of $\overline{Q}\times(-a,a)\times\mathbb{R}^{n-1}$ *and such that*

$$\begin{aligned}\mathcal{A}(\hat\rho)&=-\sum_{j,k=1}^n a_{jk}(\hat\rho,\nabla\hat\rho)\partial_j\partial_k+\sum_{j=1}^n a_j(\hat\rho,\nabla\hat\rho)\partial_j+(W\hat\rho)\partial_n,\\ \mathcal{B}(\hat\rho)&=-\sum_{j=1}^n b_j(\hat\rho,\nabla\hat\rho)\partial_j.\end{aligned}\tag{3.5}$$

Proof. Recall that $A(\rho)$ and $B(\rho)$ are just the Laplace–Beltrami operator of $(\Omega,\theta_\rho^*\eta)$ and the outer normal derivative on Γ of $(\overline{\Omega},\theta_\rho^*\eta)$, respectively, where η denotes the standard Euclidean metric on $\mathbb{R}^m$; see the proof of Lemma 2.3. Hence assertion (3.4) is obvious, since $(\mathcal{A},\mathcal{B})$ is a representation of (A,B) in local coordinates. The explicit decomposition of the coefficient of ∂_n of $\mathcal{A}$ follows from (3.2). $\square$

We close this section by determining the local representations of $K(\rho)$ and $M(\rho)$ according to the parameterization (Q,μ). In order to do this, we introduce

$$\begin{aligned}\mathcal{K}&:=\mathcal{K}(\rho):=-\partial\mathcal{A}(\hat\rho)[\cdot,\mu^*(R(\rho)b)]\in\mathcal{L}(h^{2+\alpha}(P),h^\alpha(\mathring{Q})),\\ \mathcal{M}&:=\mathcal{M}(\rho):=\partial\mathcal{B}(\hat\rho)[\cdot,\mu^*(R(\rho)b)]\in\mathcal{L}(h^{2+\alpha}(P),h^{1+\alpha}(P))\end{aligned}$$

for each $\rho\in\mathcal{V}$.

Lemma 3.2. *Given* $\rho\in\mathcal{V}$, *we have*

$$\mu^*K(\rho)=\mathcal{K}(\rho)\mu^*\qquad\textit{and}\qquad\mu^*M(\rho)=\mathcal{M}(\rho)\mu^*.$$

Proof. Fix $\rho\in\mathcal{V}$. To shorten our notation, we write $v:=R(\rho)b$ and $\hat h:=\mu^*h$ for $h\in h^{2+\alpha}(\Gamma)$. Then we have

$$\begin{aligned}\mu^*K(\rho)h=\mu^*\partial A(\rho)[h,v]&=\mu^*A(\rho+h)v-\mu^*A(\rho)v+o(h)\\ &=\mathcal{A}(\hat\rho+\hat h)\mu^*v-\mathcal{A}(\hat\rho)\mu^*v+o(\hat h)\\ &=\partial\mathcal{A}(\hat\rho)[\hat h,\mu^*v]\\ &=\mathcal{K}(\rho)\mu^*h\end{aligned}$$

as $h\to0$ in $h^{2+\alpha}(\Gamma\cap U_l)$. The second assertion can be proved analogously. $\square$

Lemma 3.3. *There exist*

$$\widetilde{k}_j,\ \widetilde{m}_j\in C^\infty(Q\times(-a,a)\times\mathbb{R}^{n-1},\mathbb{R}),\quad j=0,\dots,n-1,$$

such that

$$\begin{aligned}\mathcal{K}h&=-\partial_n[\mu^*(R(\rho)b)]Wh+\sum_{j=1}^{n-1}k_j(\hat\rho,\nabla\hat\rho)\partial_jh+k_0(\hat\rho,\nabla\hat\rho)h,\\ \mathcal{M}h&=\sum_{j=1}^{n-1}m_j(\hat\rho,\nabla\hat\rho)\partial_jh+m_0(\hat\rho,\nabla\hat\rho)h\end{aligned}$$

for each $h \in h^{2+\alpha}(P)$. Here again, k_j and m_j denote the Nemitskii operators induced by $\widetilde{k}_j$ and $\widetilde{m}_j$, respectively.

Proof. The above assertions follow easily from Lemma 3.1. □

4. Fourier multiplier operators. In this section we are concerned with linear differential operators having constant coefficients, obtained by freezing the local representation $(\mathcal{A}, \mathcal{B})$ of (A, B) at $\rho \in \mathcal{V}$ and at $0 \in Q$. These operators are used to associate a Fourier multiplier operator $\mathcal{G}_1$ to the Fréchet derivative $\partial\Phi(\rho)$ of Φ at ρ.

Throughout this section we fix $\rho \in \mathcal{V}$ and $l \in \{1, \ldots, m_\kappa\}$. Of course, all operators appearing in this section will depend on the choice (ρ, l). However, we will suppress this dependence throughout this section. Let $\mathrm{H}^n = \mathbb{R}^{n-1} \times (0,1)$ denote the truncated half-space in $\mathbb{R}^n$, and let γ_0 denote the restriction operator from H^n to $\mathbb{R}^{n-1} \times \{0\} \equiv \mathbb{R}^{n-1}$. Moreover, we set

$$a^0_{jk} := a_{jk}(\hat{\rho})(0), \qquad b^0_j := b_j(\hat{\rho})(0), \qquad 1 \le j,\ k \le n, \tag{4.1}$$

and we define the following linear differential operators with constant coefficients:

$$\mathcal{A}_0 := -\sum_{j,k=1}^{n} a^0_{jk}\partial_j\partial_k, \qquad \mathcal{B}_0 := -\sum_{j=1}^{n} b^0_j \gamma_0 \partial_j.$$

Furthermore, let

$$\vec{a} := (a^0_{1n}, \ldots, a^0_{(n-1)n}), \qquad a_0 := \sum_{j,k=1}^{n-1} a^0_{jk}\xi^j\xi^k, \quad \xi \in \mathbb{R}^{n-1},$$

and define for fixed $\xi \in \mathbb{R}^{n-1}$ the following parameter-dependent quadratic polynomial:

$$q_\xi(z) := 1 + a_0(\xi) + 2i(\vec{a}|\xi)z - a^0_{nn}z^2, \quad z \in \mathbb{C}.$$

Since the matrix $[a^0_{jk}]$ is positive definite, it follows that, given $\xi \in \mathbb{R}^{n-1}$, there exists exactly one root $\lambda(\xi)$ of $q_\xi(\cdot)$ with positive real part, which is given by

$$\lambda(\xi) = \frac{i(\vec{a}|\xi)}{a^0_{nn}} + \frac{1}{a^0_{nn}}\sqrt{a^0_{nn}(1 + a_0(\xi)) - (\vec{a}|\xi)^2}.$$

Finally, we set

$$\vec{b} := (b^0_1, \ldots, b^0_{n-1}), \qquad \vec{m} := (m^0_1, \ldots, m^0_{n-1}).$$

In the following, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and the inverse Fourier transform, respectively, in $\mathbb{R}^{n-1}$. We are now ready to introduce the following Fourier multiplier operators, acting on functions defined on $\mathbb{R}^{n-1}$.

$$\mathcal{T}_0 g(x,y) := [\mathcal{F}^{-1}e^{-\lambda(\cdot)y}\mathcal{F}g](x), \tag{4.2}$$

where $g \in h^{2+\alpha}(\mathbb{R}^{n-1})$ and $(x,y) \in \mathrm{H}^n$. Moreover,

$$\mathcal{S}_0 h(x,y) := \left[\mathcal{F}^{-1}(1 - e^{-\lambda(\cdot)y})\frac{1}{1 + a_0(\cdot)}\,\mathcal{F}h\right](x), \tag{4.3}$$

for $h \in h^{\alpha}(\mathbb{R}^{n-1})$ and $(x,y) \in \mathrm{H}^n$. Then it can be shown that

$$\begin{aligned} &\mathcal{T}_0 \in \mathcal{L}(h^{2+\alpha}(\mathbb{R}^{n-1}), h^{2+\alpha}(\mathrm{H}^n)),\\ &\mathcal{S}_0 \in \mathcal{L}(h^{\alpha}(\mathbb{R}^{n-1}), h^{2+\alpha}(\mathrm{H}^n)); \end{aligned} \tag{4.4}$$

see Appendices A and B in [12]. Next note that the function $u = \mathcal{T}_0 g$ solves the elliptic boundary value problem

$$(1+\mathcal{A}_0)u = 0 \quad \text{in} \quad \mathrm{H}^n, \qquad \gamma_0 u = g \quad \text{on} \quad \mathbb{R}^{n-1},$$

whereas $v = \mathcal{S}_0 h$ is a solution of

$$(1+\mathcal{A}_0)v = h \quad \text{in} \quad \mathrm{H}^n, \qquad \gamma_0 v = 0 \quad \text{on} \quad \mathbb{R}^{n-1},$$

where we use the same notation for the extended function $\tilde{h}(x,y) := h(x)$, $(x,y) \in \mathrm{H}^n = \mathbb{R}^{n-1} \times (0,1)$. In addition, we define

$$\begin{aligned} &k^0 := \big(\partial_n[\mu^*(R(\rho)b)]\big)(0),\\ &w_0^{jk} := w^{jk}(0), \qquad 1 \le j,k \le n-1. \end{aligned} \tag{4.5}$$

Note that $\big(\partial_n[\mu^*(R(\rho)b)]\big)(0) = -\big(\partial_\nu[R(\rho)b]\big)(\mu(0))$. Hence, it follows from (2.3) that k^0 is positive. Given $h \in h^{2+\alpha}(\mathbb{R}^{n-1})$, let

$$(\mathcal{K}_0 h)(x) := -k^0 \left[1 - \sum_{j,k=1}^{n-1} w_0^{jk} \partial_j \partial_k\right] h(x), \quad x \in \mathbb{R}^{n-1}. \tag{4.6}$$

It is then obvious that

$$\mathcal{K}_0 \in \mathcal{L}(h^{2+\alpha}(\mathbb{R}^{n-1}), h^{\alpha}(\mathbb{R}^{n-1})). \tag{4.7}$$

Similarly, we set $m_j^0 := m_j(\hat{\rho})(0)$ and define

$$\mathcal{M}_0 h := \sum_{j=1}^{n-1} m_j^0 \partial_j h, \qquad h \in h^{2+\alpha}(\mathbb{R}^{n-1}).$$

Now let $t \in [0,1]$ be given and set

$$\mathcal{G}_t := t(\mathcal{B}_0 \mathcal{S}_0 \mathcal{K}_0 + \mathcal{M}_0) + (1-t)\mathcal{B}_0 \mathcal{T}_0.$$

Observe that $\mathcal{G}_t \in \mathcal{L}(h^{2+\alpha}(\mathbb{R}^{n-1}), h^{1+\alpha}(\mathbb{R}^{n-1}))$ for $t \in [0,1]$, as (4.4) and (4.7) show. Since $\mathcal{K}_0$ and $\mathcal{M}_0$ are the principal parts of $\mathcal{K}$ and $\mathcal{M}$, respectively, with coefficients fixed at $\rho \in \mathcal{V}$ and at $0 \in Q$, the operator $\mathcal{G}_1$ may be considered as the constant coefficient operator of the principal part of $\partial\Phi(\rho)$. The operator BT is called the *Dirichlet–Neumann* operator. Hence $\mathcal{G}_0$ is the constant coefficient version of the localization $\mathcal{BT}$ of BT; see also [11]. We should mention that we slightly modified the concepts and notations as introduced in [11] and [12]. However, an inspection of the proofs given in [12] show that formula (4.4) can be proved in the same way by using Fourier multiplier results in Hölder spaces; see [12, App. A.]. We can now prove the following result.

LEMMA 4.1. *Given $t \in [0,1]$, the operator $\mathcal{G}_t$ is a Fourier multiplier operator with symbol g_t; i.e., $\mathcal{G}_t = \mathcal{F}^{-1} g_t \mathcal{F}$ where*

$$g_t(\xi) := b_n^0 \lambda(\xi) \left\{ (1-t) + t k^0 \, \frac{(1 + w_0^{jk} \xi_j \xi_k)}{1 + a_0(\xi)} \right\} + i \big\{ ((t-1)\vec{b} + t\vec{m} | \xi) \big\}$$

for all $\xi \in \mathbb{R}^{n-1}$.

Proof. (a) In a first step we provide a representation of $\mathcal{S}_0 \mathcal{K}_0$. It is an immediate consequence of (4.6) that the Fourier transform of $\mathcal{K}_0 h$ is given by

$$(\mathcal{F}\mathcal{K}_0 h)(\xi) = -k^0 \, (1 + w_0^{jk} \xi_j \xi_k)(\mathcal{F}h)(\xi)$$

for $h \in h^{2+\alpha}(\mathbb{R}^{n-1})$ and $\xi \in \mathbb{R}^{n-1}$. Now it follows from (4.3) that

$$(\mathcal{F}S_0\mathcal{K}_0 h)(\xi, y) = -(1 - e^{-\lambda(\xi) y}) k^0 \, \frac{(1 + w_0^{jk} \xi_j \xi_k)}{1 + a_0(\xi)} (\mathcal{F}h)(\xi), \tag{4.8}$$

where $\xi \in \mathbb{R}^{n-1}$ and $y \in (0,1)$.

(b) Observe that $\gamma_0 \partial_j u = \partial_j \gamma_0 u$ for $u \in h^{2+\alpha}(\mathrm{H}^n)$ and $j = 1, \ldots, n-1$. Hence (4.8) yields

$$\mathcal{B}_0 \mathcal{S}_0 \mathcal{K}_0 h = \mathcal{F}^{-1} \left[b_n^0 \lambda(\xi) k^0 \, \frac{(1 + w_0^{jk} \xi_j \xi_k)}{1 + a_0(\xi)} \mathcal{F}h \right]. \tag{4.9}$$

From formula (4.2) we infer that

$$b_j^0 \gamma_0 \partial_j \mathcal{T}_0 = \mathcal{F}^{-1}[\xi \mapsto i b_j^0 \xi_j] \mathcal{F}, \qquad j = 1, \ldots, n-1,$$

and

$$b_n^0 \gamma_0 \partial_n \mathcal{T}_0 = -\mathcal{F}^{-1} b_n^0 \lambda(\cdot) \mathcal{F}.$$

Hence we find that

$$\mathcal{B}_0 \mathcal{T}_0 = \mathcal{F}^{-1}[\xi \mapsto b_n^0 \lambda(\xi) - i(\vec{b}|\xi)] \mathcal{F}. \tag{4.10}$$

Finally, it is clear that

$$\mathcal{M}_0 = \mathcal{F}^{-1}[\xi \mapsto i(\vec{m}|\xi)] \mathcal{F}. \tag{4.11}$$

Combining (4.9)–(4.11), we get the assertion. □

As a first consequence of Lemma 4.1, we show that $-\mathcal{G}_t$ generates for each $t \in [0,1]$ a strongly continuous analytic semigroup on $h^{1+\alpha}(\mathbb{R}^{n-1})$. To make this precise we need a few definitions. To begin with, assume that $\alpha_* > 0$, $\sigma > 0$ and let

$$\begin{aligned} \mathcal{E}ll\mathcal{S}_\sigma^\infty(\alpha_*) := \big\{ & a \in C^\infty(\mathbb{R}^{n-1} \times (0,\infty)) \,;\, a \text{ is positively homogeneous} \\ & \text{of degree } \sigma, \text{ all derivatives of } a \text{ are bounded on } |\xi|^2 + \mu^2 = 1, \\ & \text{and } \operatorname{Re} a(\xi,\mu) \geq \alpha_* (|\xi|^2 + \mu^2)^{\sigma/2},\ (\xi,\mu) \in \mathbb{R}^{n-1} \times (0,\infty) \big\}. \end{aligned}$$

Given two Banach spaces E_0 and E_1 such that E_1 is continuously and densely embedded in E_0, let $\mathcal{H}(E_1, E_0)$ denote the set of all $A \in \mathcal{L}(E_1, E_0)$ such that $-A$, considered as an unbounded operator in E_0, generates a strongly continuous analytic semigroup on E_0. It is known (see Remark I.1.2.1(a) in [2]) that $A \in \mathcal{L}(E_1, E_0)$ belongs to $\mathcal{H}(E_1, E_0)$ if there exist positive constants C and λ_* such that

$$\begin{aligned} &\lambda_* + A \in \mathrm{Isom}(E_1, E_0), \\ &|\lambda|\,\|x\|_{E_0} + \|x\|_{E_1} \leq C\|(\lambda + A)x\|_{E_0}, \quad x \in E_1,\ \lambda \in [\mathrm{Re}\, z \geq \lambda_*]. \end{aligned} \tag{4.12}$$

THEOREM 4.2. *Suppose that* (2.3) *holds. Then*

$$\mathcal{G}_t \in \mathcal{H}(h^{2+\alpha}(\mathbb{R}^{n-1}), h^{1+\alpha}(\mathbb{R}^{n-1})), \qquad t \in [0,1].$$

Proof. (a) Basically, the idea is to use Lemma 4.1 together with appropriate results on Fourier multipliers to verify the generation property of $\mathcal{G}_t$. Having this intention, it is well known that homogeneous symbols are much easier to handle. Hence, in a first step we introduce a parameter-dependent version of the symbol g_t, which is positively homogeneous of degree 1. Given $(\xi, \mu) \in \mathbb{R}^{n-1} \times (0, \infty)$, let

$$\lambda(\xi, \mu) := \frac{i(\vec{a}|\xi)}{a_{nn}^0} + \frac{1}{a_{nn}^0}\sqrt{a_{nn}^0(\mu^2 + a_0(\xi)) - (\vec{a}|\xi)^2}$$

and $r(\xi, \mu) := \mathrm{Re}(\lambda(\xi, \mu))$. Then we set

$$\widetilde{g}_t(\xi, \mu) := b_n^0 \lambda(\xi, \mu)\left\{(1-t) + tk^0\,\frac{(\mu^2 + w_0^{jk}\xi_j\xi_k)}{\mu^2 + a_0(\xi)}\right\} + i\{((t-1)\vec{b} + t\vec{m}|\xi)\},$$

for $(\xi, \mu) \in \mathbb{R}^{n-1} \times (0, \infty)$ and $t \in [0,1]$. Obviously, $\widetilde{g}_t(\cdot, 1) = g_t$. Moreover, it is clear that $\widetilde{g}_t \in C^\infty(\mathbb{R}^{n-1} \times (0,\infty), \mathbb{C})$ and that each $\widetilde{g}_t$ is positively homogeneous of degree 1. In addition, it is easily verified that all derivatives of a are bounded on $|\xi|^2 + \mu^2 = 1$.

(b) Observe that $k^0 > 0$, thanks to assumption (2.4) and (2.3). In addition, we know from (3.4) and (3.5) that $a_{nn}^0 > 0$ and $b_n^0 > 0$. Furthermore, there exist positive constants K and r_* such that

$$\mu^2 + a_0(\xi) \leq K(\mu^2 + |\xi|^2), \qquad r(\xi, \mu) \geq r_*\sqrt{\mu^2 + |\xi|^2}$$

for all $(\xi, \mu) \in \mathbb{R}^{n-1} \times (0, \infty)$. The first estimate follows immediately from the definition of a_0. The second one is a consequence of the ellipticity of $[a_{jk}]_{1 \leq j,k \leq n}$. Finally, recall that w is uniformly positive definite; see section 3. Hence there is a positive constant $w_* > 0$ such that $(\mu^2 + w_0^{jk}\xi_j\xi_k) \geq w_*(\mu^2 + |\xi|^2)$ for all $(\xi, \mu) \in \mathbb{R}^{n-1} \times (0, \infty)$. This leads to an estimate

$$\begin{aligned} \mathrm{Re}\,\widetilde{g}_t(\xi, \mu) =&\ b_n^0 r(\xi, \mu)\left\{(1-t) + tk^0\,\frac{(\mu^2 + w_0^{jk}\xi_j\xi_k)}{\mu^2 + a_0(\xi)}\right\} \\ \geq&\ b_n^0 r_*\sqrt{\mu^2 + |\xi|^2}\left\{(1-t) + tk^0\,\frac{w_*(\mu^2 + |\xi|^2)}{K(\mu^2 + |\xi|^2)}\right\} \\ \geq&\ b_n^0 r_*\sqrt{\mu^2 + |\xi|^2}\{(1-t) + tk_*\}, \end{aligned}$$

where $k_* := k^0 K^{-1} w_* > 0$. Now, letting

$$\alpha_* := r_* b_n^0 \min\{1, k_*\} > 0,$$

we find that $\widetilde{g}_t \in \mathcal{E}ll\mathcal{S}_1^\infty(\alpha_*)$ for all $t \in [0,1]$. Now the assertion is implied by a general result due to Amann, which in particular states that given $a \in \mathcal{E}ll\mathcal{S}_1^\infty(\alpha_*)$ and $\mu_0 > 0$; it follows that $a(\cdot, \mu_0) \in \mathcal{H}(h^{2+\alpha}(\mathbb{R}^{n-1}), h^{1+\alpha}(\mathbb{R}^{n-1}))$; see [3]. □

5. Perturbations. In this section we prove that, given $\rho \in \mathcal{V}$, the linearization $-\partial\Phi(\rho)$ of $-\Phi$ at ρ generates a strongly continuous analytic semigroup on $h^{1+\alpha}(\Gamma)$. The main technical tool is a perturbation result contained in Lemma 5.1. To state this result we need some preparation. First let

$$\partial\Phi_t(\rho) := t\partial\Phi(\rho) + (1-t)B(\rho)T(\rho)$$

for $\rho \in \mathcal{V}$ and $t \in [0,1]$. Obviously, $\partial\Phi_t(\rho)$ is a convex combination connecting $\partial\Phi(\rho)$ and the Dirichlet–Neumann operator $B(\rho)T(\rho)$; see [11].

Next, given $\kappa \in (0,a]$, choose smooth test functions $\psi_l \in \mathcal{D}(U_l)$ such that $\{(U_l,\psi_l)\ ;\ 1 \le l \le m_\kappa\}$ is a partition of unity on $\mathcal{R}_\kappa$; see section 3 for the definition of $\mathcal{R}_\kappa$. Call such a family $\{(U_l,\psi_l)\ ;\ 1 \le l \le m_\kappa\}$ a (finite) *localization sequence* for $\mathcal{R}_\kappa$. Moreover, we fix $\hat{x}_l \in \Gamma$ such that $\hat{x}_l \in U_l$, $l = 1,\dots,m_\kappa$. We may further assume that $\mu_l(0) = \hat{x}_l$ for $l = 1,\dots,m_\kappa$.

To economize our notation, the symbols $|\cdot|_s$ and $\|\cdot\|_s$ are exclusively used for the norms in $h^s(\mathbb{R}^{n-1})$ and $h^s(\Gamma)$, respectively.

Finally, throughout this section we fix $\rho \in \mathcal{V}$ and $\beta \in (0,\alpha)$.

LEMMA 5.1. *Given $\varepsilon > 0$, there exists $\kappa \in (0,a]$, a localization sequence $\{(U_l,\psi_l);$ $1 \le l \le m_\kappa\}$ for $\mathcal{R}_\kappa$, and a positive constant $C := C(\rho,\varepsilon,\kappa)$ such that*

$$|\mu_l^*(\psi_l\partial\Phi_t(\rho)h) - \mathcal{G}_t(\rho,l)\mu_l^*(\psi_l h)|_{1+\alpha} \le \varepsilon|\mu_l^*(\psi_l h)|_{2+\alpha} + C\|h\|_{2+\beta}$$

for all $h \in h^{2+\alpha}(\Gamma)$, $l \in \{1,\dots,m_\kappa\}$, and $t \in [0,1]$.

Proof. (a) We fix $\rho \in \mathcal{V}$, $l \in \{1,\dots,m_\kappa\}$ and suppress the pair (ρ,l) in our notation. Moreover, given $\varepsilon > 0$ and $\beta \in (0,\alpha)$, we only show explicitly the existence of a positive constant C such that

$$|\mu^*(\psi BSKh) - \mathcal{B}_0\mathcal{S}_0\mathcal{K}_0\mu^*(\psi h)|_{1+\alpha} \le \varepsilon|\mu^*(\psi h)|_{2+\alpha} + C\|h\|_{2+\beta}$$

for all $h \in h^{2+\alpha}(\Gamma)$. The remaining two terms

$$|\mu^*(\psi BTh) - \mathcal{B}_0\mathcal{T}_0\mu^*(\psi h)|_{1+\alpha}, \qquad |\mu^*(\psi Mh) - \mathcal{M}_0\mu^*(\psi h)|_{1+\alpha}$$

can be estimated similarly (and are even easier to handle). Our argumentation follows the proof of Lemma 6.1 in [12] and uses in particular obvious generalizations of Lemmas 6.5, 6.6, and 6.7 in [12] to the n-dimensional case.

(b) Choose a smooth test-function $\chi \in \mathcal{D}(U)$ such that $\chi|\mathrm{supp}(\psi) = 1$. Then we have

$$\mu^*\psi BSK - \mathcal{B}_0\mathcal{S}_0\mathcal{K}_0\mu^*\psi = \mu^*\chi BSK\psi - \mathcal{B}_0\mathcal{S}_0\mathcal{K}_0\mu^*\chi\psi - \mu^*\chi[BSK,\psi],$$

where ψ and χ also denote the linear operators induced by pointwise multiplication by ψ and χ, respectively, and where $[A,B] := AB - BA$ denotes the commutator of A and B. It follows, essentially from Leibniz' rule (see Lemma 6.5(b) in [12]), that there exists a positive constant C such that

$$\|[BSK,\psi]h\|_{1+\alpha} \le C\|h\|_{2+\beta}, \qquad h \in h^{2+\alpha}(\Gamma).$$

Hence, it suffices to estimate the operator

$$\mu^*\chi BSK - \mathcal{B}_0\mathcal{S}_0\mathcal{K}_0\mu^*\chi.$$

In addition, we split that operator in the following way:

$$\mu^*\chi BSK - \mathcal{B}_0\mathcal{S}_0\mathcal{K}_0\mu^*\chi = \mu^*\chi BSK - \mathcal{B}_0\mu^*\chi SK + \mathcal{B}_0\{\mu^*\chi S - \mathcal{S}_0\mu^*\chi\}K + \mathcal{B}_0\mathcal{S}_0\{\mu^*\chi K - \mathcal{K}_0\mu^*\chi\}. \tag{5.1}$$

(c) Let us start with the first term $\mu^*\chi BSK - \mathcal{B}_0\mu^*\chi SK$. Again, by Leibniz' rule, the commutator $[\mu^*\chi, \mathcal{B}_0]$ can be estimated as

$$|[\mu^*\chi, \mathcal{B}_0]u|_{1+\alpha} \le C|u|_{1+\alpha,H^n}, \qquad u \in h^{2+\alpha}(\mathrm{H}^n). \tag{5.2}$$

Thus we are left to control the operator $\mu^*\chi B - (\mu^*\chi)\mathcal{B}_0\mu^*$. By the definition of $\mathcal{B}$ we get the formula

$$\mu^*\chi B - (\mu^*\chi)\mathcal{B}_0\mu^* = (\mu^*\chi)\{\mathcal{B} - \mathcal{B}_0\}\mu^*. \tag{5.3}$$

But, as in [12, Lemma 6.7(a)], we find positive constants C and C_κ such that

$$\begin{aligned} |(\mu^*\chi)\{1 + \mathcal{A}_0 - \mathcal{A}\}(\mu^* v)|_{\alpha,H^n} &+ |(\mu^*\chi)\{\mathcal{B} - \mathcal{B}_0\}(\mu^* v)|_{1+\alpha} \\ &\le C\kappa^{1-\alpha}\|v\|_{2+\alpha,\Omega} + C_\kappa\|v\|_{1+\alpha,\Omega} \end{aligned} \tag{5.4}$$

for all $v \in h^{2+\alpha}(\Omega)$. Finally, observe that

$$S \in \mathcal{L}(h^\gamma(\Omega), h^{2+\gamma}(\Omega)), \qquad K \in \mathcal{L}(h^{2+\gamma}(\Gamma), h^\gamma(\Omega)) \tag{5.5}$$

for $\gamma \in [\beta, \alpha]$ and that

$$\mu^* \in \mathrm{Diff}^\infty(h^{2+\alpha}(\Gamma \cap U), h^{2+\alpha}(P)). \tag{5.6}$$

Combining (5.2)–(5.6), we can find a $\kappa_1 \in (0, a]$ and a positive constant C such that

$$|\mu^*(\chi BSKg) - \mathcal{B}_0\mu^*(\chi SKg)|_{1+\alpha} \le \frac{\varepsilon}{3}|\mu^* g|_{2+\alpha} + C\|g\|_{2+\beta} \tag{5.7}$$

for all $g \in h^{2+\alpha}(\Gamma \cap U)$.

(d) In a next step we estimate the operator $\mu^*\chi S - \mathcal{S}_0\mu^*\chi$. To achieve this, we use the representation

$$\mu^*\chi S - \mathcal{S}_0\mu^*\chi = \mathcal{S}_0\big\{[\mathcal{A}_0, \mu^*\chi]\mu^* S + (\mu^*\chi)\{1 + \mathcal{A}_0 - \mathcal{A}\}\mu^* S\big\}, \tag{5.8}$$

which follows from Lemma 6.6 in [12]. Again, the operator $[\mathcal{A}_0, \mu^*\chi]$ is of lower order in the sense that there exists a positive constant C such that

$$|[\mathcal{A}_0, \mu^*\chi]u|_{\alpha,\mathring{Q}} \le C|u|_{1+\alpha,\mathring{Q}}, \qquad u \in h^{1+\alpha}(\mathring{Q}). \tag{5.9}$$

Hence, it follows from (5.4), (5.5), (5.6), (5.8), and (5.9) that there is a $\kappa_2 \in (0, a]$ such that

$$|\mathcal{B}_0\big\{\mu^*(\chi SKg) - \mathcal{S}_0\mu^*(\chi Kg)\big\}|_{1+\alpha} \le \frac{\varepsilon}{3}|\mu^* g|_{2+\alpha} + C\|g\|_{2+\beta} \tag{5.10}$$

for all $g \in h^{2+\alpha}(\Gamma \cap U)$.

(e) From Lemma 3.2 we know that

$$\mu^*\chi K - \mathcal{K}_0\mu^*\chi = (\mu^*\chi)\{\mathcal{K} - \mathcal{K}_0\}\mu^* + [\mu^*\chi, \mathcal{K}_0]\mu^*.$$

But here again, it follows from Leibniz' rule that there is a $C > 0$ such that

$$|[\mu^*\chi, \mathcal{K}_0]\mu^* g|_{\alpha,H^n} \leq C\|g\|_{1+\alpha}, \qquad g \in h^{2+\alpha}(\Gamma \cap U). \tag{5.11}$$

Finally, we infer from Lemma 6.7(b) in [12] that there are positive constants C and C_κ such that

$$|(\mu^*\chi)\{\mathcal{K} - \mathcal{K}_0\}g|_{\alpha,H^n} \leq \kappa^{1-\alpha}C|\mu^* g|_{2+\alpha} + C_\kappa\|g\|_{1+\alpha} \tag{5.12}$$

for all $g \in h^{2+\alpha}(\Gamma \cap U)$. Since $\mathcal{B}_0\mathcal{S}_0 \in \mathcal{L}(h^\alpha(\mathrm{H}^n), h^{1+\alpha}(\Gamma))$, we conclude from (5.11) and (5.12) that there is a $\kappa_3 \in (0, a]$ and a $C > 0$ such that

$$|\mathcal{B}_0\mathcal{S}_0\big\{\mu^*\chi K - \mathcal{K}_0\mu^*\chi\big\}g|_{1+\alpha} \leq \frac{\varepsilon}{3}|\mu^* g|_{2+\alpha} + C\|g\|_{2+\beta} \tag{5.13}$$

for all $g \in h^{2+\alpha}(\Gamma \cap U)$. Now, letting $\kappa := \min\{\kappa_1, \kappa_2, \kappa_3\}$, the assertion follows from (5.7), (5.10), and (5.13). □

THEOREM 5.2. *We have*

$$\partial\Phi_t(\rho) \in \mathcal{H}(h^{2+\alpha}(\Gamma), h^{1+\alpha}(\Gamma)), \qquad \rho \in \mathcal{V}, \quad t \in [0,1].$$

Proof. (a) In a first step we provide a parameter-dependent a priori estimate for $\partial\Phi_t(\rho)$. To begin with, we know from Theorem 4.2 that there are positive constants λ_1 and C_1, independent of $\kappa \in (0, a]$ and $l \in \{1, \dots, m_\kappa\}$, such that

$$|g|_{2+\alpha} + |\lambda||g|_{1+\alpha} \leq C_1|(\lambda + \mathcal{G}_t(\rho, l))g|_{1+\alpha} \tag{5.14}$$

for all $g \in h^{2+\alpha}(\mathbb{R}^{n-1})$, $\lambda \in [\operatorname{Re} z \geq \lambda_1]$, and $l \in \{1, \dots, m_\kappa\}$. Furthermore, Lemma 5.1 guarantees the existence of positive constants κ, C_2, and a localization sequence $\{(U_l, \psi_l)\,;\, 1 \leq l \leq m_\kappa\}$ such that

$$|\mu_l^*\big(\psi_l\partial\Phi_t(\rho)h\big) - \mathcal{G}_t(\rho, l)\mu_l^*(\psi_l h)|_{1+\alpha} \leq \frac{1}{2C_1}|\mu_l^*(\psi_l h)|_{2+\alpha} + C_2\|h\|_{2+\beta}$$

for all $h \in h^{2+\alpha}(\Gamma)$, $l \in \{1, \dots, m_\kappa\}$, and $t \in [0,1]$. Consequently, it follows from (5.14) that

$$\begin{aligned} |\mu_l^*(\psi_l h)|_{2+\alpha} &+ |\lambda||\mu_l^*(\psi_l h)|_{1+\alpha} \\ &\leq 2C_1\big\{|\mu_l^*\big(\psi_l(\lambda + \partial\Phi_t(\rho))h\big)|_{1+\alpha} + C_2\|h\|_{2+\beta}\big\} \end{aligned} \tag{5.15}$$

for all $h \in h^{2+\alpha}(\Gamma)$, $\lambda \in [\operatorname{Re} z \geq \lambda_1]$, $l \in \{1, \dots, m_\kappa\}$, and $t \in [0,1]$. Next observe that

$$\big[h \mapsto \max_{1 \leq l \leq m_\kappa} |\mu_l^*(\psi_l h)|_{k+\alpha}\big]$$

defines an equivalent norm on $h^{k+\alpha}(\Gamma)$, $k = 1, 2$, due to the fact that the family $\{(U_l, \psi_l)\,;\, 1 \leq l \leq m_\kappa\}$ is a localization sequence for $\mathcal{R}_\kappa$; see [24]. Hence (5.15) implies the existence of a positive constant C such that

$$\|h\|_{2+\alpha} + |\lambda|\|h\|_{1+\alpha} \leq \frac{C}{2}\|(\lambda + \partial\Phi_t(\rho))h\|_{1+\alpha} + C\|h\|_{2+\beta} \tag{5.16}$$

for all $h \in h^{2+\alpha}(\Gamma)$, $\lambda \in [\operatorname{Re} z \geq \lambda_1]$, and $t \in [0,1]$.

Finally, let $(\cdot,\cdot)^0_{\theta,\infty}$ denote the continuous interpolation functor of Da Prato and Grisvard; see [6]. It is known that

$$h^{2+\beta}(\Gamma) = \big(h^{1+\alpha}(\Gamma), h^{2+\alpha}(\Gamma)\big)^0_{1-\alpha+\beta,\infty}. \tag{5.17}$$

Hence there exists a positive constant C_3 such that

$$\|h\|_{2+\beta} \le \frac{1}{2C}\|h\|_{2+\alpha} + C_3\|h\|_{1+\alpha}, \qquad h \in h^{2+\alpha}(\Gamma).$$

Now we conclude from (5.17) that

$$\|h\|_{2+\alpha} + |\lambda|\|h\|_{1+\alpha} \le C\|(\lambda + \partial\Phi_t(\rho))h\|_{1+\alpha} \tag{5.18}$$

for all $h \in h^{2+\alpha}(\Gamma)$, $\lambda \in [\operatorname{Re} z \ge \lambda_*]$, and $t \in [0,1]$, where we have set $\lambda_* := 2\max\{\lambda_1, CC_3\}$.

(b) In view of (4.12) and (5.18), it remains to prove that $\partial\Phi_t(\rho)$ is surjective for each $t \in [0,1]$. Moreover, since the estimate (5.18) is uniform in $t \in [0,1]$, a well-known homotopy argument (see Theorem 5.2 in [17]) implies that it is sufficient to prove that $\partial\Phi_0(\rho)$ is onto. Thus, let $g \in h^{1+\alpha}(\Gamma)$ be given. Then we find, as in the proof of Lemma 2.4, a unique $v \in h^{2+\alpha}(\Omega)$ such that

$$\big(A(\rho), \lambda_*\gamma + B(\rho), C\big)v = (0, g, 0). \tag{5.19}$$

The first and the third components of this identity imply that

$$T(\rho)\gamma v = \big(A(\rho), \gamma, B(\rho)\big)^{-1}(0, \gamma v, 0) = v;$$

see section 2 for the definition of the operator $T(\rho)$. Now, putting $h := \gamma v \in h^{2+\alpha}(\Gamma)$, the second component of (5.19) gives

$$\big(\lambda_* + B(\rho)T(\rho)\big)h = \big(\lambda_*\gamma + B(\rho)\big)v = g,$$

which completes our argumentation. □

Remark 5.3. Let $\rho \in \mathcal{V}$ be given. Then the proofs of Theorems 4.2 and 5.2 show that $-\partial\Phi(\rho)$ does not generate a strongly continuous semigroup on $h^{1+\alpha}(\Gamma)$ if $b \in h^{2+\alpha-\delta}(J)\backslash\{0\}$ is nonpositive. Hence, for such b, the linearized evolution equation for the moving boundary

$$\partial_t\sigma + \partial\Phi(\rho)\sigma = 0, \qquad \sigma(0) = \sigma_0$$

is not well posed in $h^{1+\alpha}(\Gamma)$ in the sense of Hadamard.

Proof of Theorem 1. Let $\rho_0 \in \mathcal{V}$ be given. Thanks to Lemmas 2.2 and 2.5 we only have to prove the existence and uniqueness of a classical Hölder solution of (2.5). To show this, fix $\beta \in (0,\alpha)$. Then it follows from Theorem 5.2 that

$$\partial\Phi(\rho) \in \mathcal{H}(h^{2+\gamma}(\Gamma), h^{1+\gamma}(\Gamma)), \quad \rho \in \mathcal{V},\ \gamma \in [\beta,\alpha].$$

From this and the known fact that little Hölder spaces are stable under continuous interpolation one finds that

$$\partial\Phi(\rho) \in \mathcal{M}_1(h^{2+\alpha}(\Gamma), h^{1+\alpha}(\Gamma)), \qquad \rho \in \mathcal{V}, \tag{5.20}$$

where $\mathcal{M}_1(E_1, E_0)$ denotes the class of all operators in $\mathcal{L}(E_1, E_0)$, having the property of maximal regularity in the sense of Da Prato and Grisvard [6]; see also [4] and [23]. The assertions now follow from Theorem 2.7 in [4].

Acknowledgments. The authors would like to thank Sigurd Angenent for helpful suggestions leading to an improvement of preliminary results. In addition, we thank the referee for carefully reading the manuscript and for suggesting some improvements.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] H. Amann, *Linear and Quasilinear Parabolic Problems* I, Birkhäuser, Basel, 1995.

[3] H. Amann, *Linear and Quasilinear Parabolic Problems* II, book in preparation.

[4] S. B. Angenent, *Nonlinear analytic semiflows*, Proc. Roy. Soc. Edinburgh, 115A (1990), pp. 91–107.

[5] J. Crank, *Free and Moving Boundary Problems*, Clarendon Press, Oxford, UK, 1984.

[6] G. Da Prato and P. Grisvard, *Equations d'évolution abstraites nonlinéaires de type parabolique*, Ann. Mat. Pura Appl., 120 (1979), pp. 329–396.

[7] E. DiBenedetto and A. Friedman, *The ill-posed Hele-Shaw model and the Stefan problem for supercooled water*, Trans. Amer. Math. Soc., 282 (1984), pp. 183–204.

[8] C. M. Elliott, *On a variational inequality formulation of an electrical machining moving boundary problem and its approximation by the finite element method*, J. Inst. Math. Appl., 25 (1980), pp. 121–131.

[9] C. M. Elliott and V. Janovsky, *A variational inequality approach to Hele-Shaw flow with a moving boundary*, Proc. Roy. Soc. Edinburgh, 88A (1981), pp. 93–107.

[10] C. M. Elliot and J. R. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Pitman, Boston, 1982.

[11] J. Escher, *The Dirichlet-Neumann operator on continuous functions*, Ann. Scuola Norm. Pisa, XXI (1994), pp. 235–266.

[12] J. Escher and G. Simonett, *Maximal regularity for a free boundary problem*, Nonlinear Differential Equations Appl., 2 (1995), pp. 463–510.

[13] J. Escher and G. Simonett, *Analyticity of the interface in a free boundary problem*, Math. Ann., 305 (1996), pp. 439–459.

[14] A. Fasano and M. Primicerio, *New results on some classical parabolic free boundary problems*, Quart. Appl. Math., 38 (1981), pp. 439–460.

[15] A. Fasano and M. Primicerio, *Blow-up and regularization for the Hele-Shaw problem*, in Variational and Free Boundary Problems, IMA 53, A. Friedman and J. Spruck, eds., Springer, New York, 1993, pp. 73–85.

[16] A. Friedman, *Time dependent free boundary problems*, SIAM Rev., 21 (1979), pp. 213–221.

[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1977.

[18] E. I. Hanzawa, *Classical solutions of the Stefan problem*, Tôhoku Math. J., 33 (1981), pp. 297–335.

[19] A. Lunardi, *Interpolation spaces between domains of elliptic operators and spaces of continuous functions with applications to nonlinear parabolic equations*, Math. Nachr., 121 (1985), pp. 295–318.

[20] M. Reissig, *The existence and uniqueness of analytic solutions for a moving boundary problem for Hele-Shaw flows in the plane*, Nonlin. Anal., T.M.&A., 23 (1994), pp. 565–576.

[21] S. Richardson, *Hele-Shaw flows with a free boundary produced by the injection of the fluid into a narrow channel*, J. Fluid Mech., 56 (1972), pp. 609–618.

[22] J. Steinbach and W. Weinelt, *Mathematical modelling of the injection moulding process*, Adv. Math. Sci. Appl., 1 (1992), pp. 137–156.

[23] G. Simonett, *Quasilinear parabolic equations and semiflows*, in Evolution Equations, Control Theory, and Biomathematics, Lecture Notes in Pure and Appl. Math. 155, M. Dekker, New York, 1994, pp. 523–536.

[24] H. Triebel, *Theory of Function Spaces*, Birkhäuser, Basel, 1983.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1048–1063, September 1997

A DEGENERATE PARABOLIC-HYPERBOLIC SYSTEM MODELING THE SPREADING OF SURFACTANTS*

MICHAEL RENARDY[†]

Abstract. We consider the initial-boundary value problem for a coupled parabolic-hyperbolic system for which the parabolic part degenerates. The problem arises in studying the spreading of surfactants on thin films, and the degeneracy occurs if either the surfactant concentration or the film height vanishes. Both cases will be considered. If the film height vanishes, the front does not advance. If, on the other hand, the surfactant concentration vanishes, then solutions with stationary or moving fronts are possible.

Key words. surfactant spreading, degenerate parabolic equations, parabolic-hyperbolic systems

AMS subject classifications. 35K65, 35Q35, 76D45

PII. S0036141096299120

1. Introduction.

We consider the spreading of a surfactant on a thin fluid film. The equations modeling this problem were introduced in [3], motivated by an application involving the medical treatment of premature infants. It is assumed that the flow in the film is driven entirely by the shear stress resulting from surface tension gradients, and that the surfactant is advected with the flow. Under these assumptions, the equations of motion, in dimensionless form, are [3]

$$h_t + \operatorname{div}\left(\frac{h^2}{2}\nabla\sigma(\Gamma)\right) = 0,$$

$$\Gamma_t + \operatorname{div}\left(h\Gamma\nabla\sigma(\Gamma)\right) = 0. \tag{1}$$

Here h is the film thickness, Γ is the surfactant concentration, and σ is the concentration-dependent surface tension coefficient. For a surfactant, $\sigma'(\Gamma) < 0$.

From a mathematical point of view, equations (1) form a parabolic-hyperbolic system. Even in the absence of molecular diffusion of the surfactant, there is an effective diffusion associated with the spreading of surfactant by the self-induced flow. This diffusion, however, is present only if both h and Γ are positive, and hence the parabolic part of the equation degenerates if either h or Γ vanishes. If molecular diffusion is added, this degeneracy is removed.

In earlier work [6], we have studied the case of positive molecular diffusion, where the parabolic equation is nondegenerate. The well-posedness of initial and initial-boundary value problems was established. In addition, it was shown that shocks associated with the hyperbolic part of the equation can develop from smooth data in finite time. For this result, it is essential that molecular diffusion be present; if it is zero, then hyperbolic shocks are impossible. The fronts which will arise in this paper are fundamentally different from hyperbolic shocks, even though they also involve

*Received by the editors February 8, 1996; accepted for publication April 1, 1996.
http://www.siam.org/journals/sima/28-5/29912.html

[†]Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123 (renardym@math.vt.edu). This author was supported by National Science Foundation grant DMS-9306635 and by the Office of Naval Research grant N00014-92-J-1664. A part of this research was carried out while the author was visiting the Isaac Newton Institute, University of Cambridge.

jumps in h and Γ_x; they are associated with the parabolic part of the problem, rather than the hyperbolic part.

Throughout this paper, we shall be concerned with "smooth" solutions. Although there is an extensive literature on degenerate parabolic equations (see [4] for an overview), results guaranteeing the existence of smooth solutions seem to be known only in one space dimension [1], [2], [5]. We therefore consider only the one-dimensional case of (1):

$$h_t + \frac{1}{2}(h^2\sigma(\Gamma)_x)_x = 0,$$

$$\Gamma_t + (h\Gamma\sigma(\Gamma)_x)_x = 0. \tag{2}$$

We seek solutions for $t > 0$, on the interval $0 < x < \phi(t)$, where $\phi(t)$ is an unknown free boundary. We have initial conditions

$$\phi(0) = 1,\ h(x,0) = h_0(x),\ \Gamma(x,0) = \Gamma_0(x). \tag{3}$$

It is understood that h_0 and Γ_0 are strictly positive except at the right end point, where either $h_0(1) = 0$ or $\Gamma_0(1) = 0$. In the first case, $\phi(t) = 1$ for $t > 0$, while in the second case, $\phi(t)$ has to be determined as part of the problem, subject to the condition

$$\Gamma(\phi(t), t) = 0. \tag{4}$$

At the left end point we shall impose, for simplicity, a no-flux condition,

$$\Gamma_x(0,t) = 0. \tag{5}$$

Following the lead of [6], we make the transformation $h = \Gamma^{1/2}p$. This leads to the new set of equations

$$p_t - \frac{1}{4}\Gamma^{-1/2}\Gamma_x p^2\sigma(\Gamma)_x + \frac{1}{2}\Gamma^{1/2}pp_x\sigma(\Gamma)_x = 0,$$

$$\Gamma_t + (p\Gamma^{3/2}\sigma(\Gamma)_x)_x = 0. \tag{6}$$

The effect of the transformation is an uncoupling of the hyperbolic and parabolic parts of the equation; note that the equation for p in (6) contains no second derivatives of Γ and can be viewed as a first-order hyperbolic equation for p when Γ is given.

2. Fronts with vanishing film thickness. In this section, we consider solutions for which $\phi(t) = 1$ for all t and $h(1) = p(1) = 0$. In (6), we substitute $p(x,t) = (1-x)q(x,t)$, leading to the new set of equations:

$$q_t - \frac{1}{4}(1-x)\Gamma^{-1/2}\Gamma_x q^2\sigma(\Gamma)_x - \frac{1}{2}\Gamma^{1/2}q^2\sigma(\Gamma)_x$$

$$+\frac{1}{2}(1-x)\Gamma^{1/2}\sigma(\Gamma)_x qq_x = 0,$$

$$\Gamma_t + ((1-x)q\Gamma^{3/2}\sigma(\Gamma)_x)_x = 0. \tag{7}$$

To complement these equations, we have the initial conditions

$$\Gamma(x,0)=\Gamma_0(x), \qquad q(x,0)=q_0(x), \tag{8}$$

and we assume that Γ_0 and q_0 are smooth and strictly positive. In addition, we have the boundary condition

$$\Gamma_x(0,t)=0; \tag{9}$$

no boundary condition is needed at $x=1$.

To solve (7)–(9), we employ the iteration

$$q_t^{n+1}-\frac{1}{4}(1-x)(\Gamma^n)^{-1/2}\Gamma_x^n(q^n)^2\sigma(\Gamma^n)_x-\frac{1}{2}(\Gamma^n)^{1/2}(q^n)^2\sigma(\Gamma^n)_x$$

$$+\frac{1}{2}(1-x)(\Gamma^n)^{1/2}\sigma(\Gamma^n)_x q^n q_x^{n+1}=0,$$

$$\Gamma_t^{n+1}+((1-x)q^n(\Gamma^n)^{3/2}\sigma'(\Gamma^n)\Gamma_x^{n+1})_x=0. \tag{10}$$

To formulate results, we need to define some function spaces. Let

$$X_k=\{u\,|\,u\in H^{k-1}(0,1),\ (1-x)u^{(k)}\in L^2(0,1)\}. \tag{11}$$

Moreover, we define $\|\cdot\|_{k,p,l}$ as the norm of a function in $W^{l,p}((0,T);X_k)$. Let Z_M be the set of all (q,Γ) such that

$$q\in\bigcap_{i=1}^{k+1}W^{k+1-i,\infty}((0,T);X_i)\cap H^{k+1}((0,T),X_0),$$

$$\Gamma\in\bigcap_{i=2}^{k+2}H^{k+2-i}((0,T);X_i)\cap H^{k+1}((0,T),X_0),$$

$$\|q\|_{0,2,k+1}+\sum_{i=1}^{k+1}\|q\|_{i,\infty,k-i+1}+\|\Gamma\|_{0,2,k+1}+\sum_{i=2}^{k+2}\|\Gamma\|_{i,2,k+2-i}\le M,$$

$$q(x,0)=q_0(x),\quad \Gamma(x,0)=\Gamma_0(x),\quad \Gamma_x(0,t)=0,$$

$$\text{The time derivatives of } q \text{ and } \Gamma \text{ up to order } k$$

$$\text{satisfy the appropriate initial condition.} \tag{12}$$

By the latter condition, we mean that the initial values of time derivatives agree with those which can be derived from the equations (7) and initial conditions (8).

The solution will be constructed as a fixed point of the iteration (10) via the contraction mapping theorem. Let $\mathcal{S}$ denote the mapping $(q^n,\Gamma^n)\mapsto(q^{n+1},\Gamma^{n+1})$, which is defined by (10) in conjunction with the initial condition (8) and boundary condition (9).

LEMMA 2.1. *Assume that Γ_0 and q_0 are strictly positive and sufficiently smooth, that the function $\Gamma \mapsto \sigma(\Gamma)$ is smooth, and that $\sigma' < 0$. Moreover, assume $k \geq 2$. Finally, assume that $\Gamma_0'(0) = 0$ and that the initial values for all time derivatives of Γ up to order $k-1$ are also compatible with the boundary condition* (9). *If M is chosen sufficiently large, and T is sufficiently small relative to M, then $\mathcal{S}$ maps Z_M into itself.*

Remark. In stating the result, we have been deliberately vague about the required smoothness of the initial data. Basically, the smoothness required of the initial data is the same as is recovered by the solution. However, this involves time derivatives of q and Γ as well as the functions themselves, and the initial values of time derivatives depend on q_0 and Γ_0 in a very complicated fashion. Hence a precise smoothness condition on q_0 and Γ_0 would be very awkward to state.

We now sketch the proof of the lemma. The first equation of (10) has the simple form

$$q_t^{n+1} + g^n q_x^{n+1} + f^n = 0; \tag{13}$$

i.e., it is simply a first-order hyperbolic equation. Since g^n vanishes at both end points, no boundary conditions are required. A simple energy estimate yields that

$$\int_0^1 (q^{n+1}(x,t))^2\,dx = \int_0^1 q_0(x)^2\,dx + \frac{1}{2}\int_0^t \int_0^1 g_x^n(x,\tau)(q^{n+1}(x,\tau))^2\,dx\,d\tau$$

$$- \int_0^t \int_0^1 f^n(x,\tau) q^{n+1}(x,\tau)\,dx\,d\tau. \tag{14}$$

This yields a bound for the L^2-norm of q^{n+1}. In a similar fashion, we can obtain a bound for the L^2-norm of $(1-x)q^{n+1}$, if we first multiply (13) by $(1-x)$. We can obtain analogous bounds for spatial and temporal derivatives of q^{n+1} by taking derivatives of (12); note that the initial values of any derivative of q^{n+1} or Γ^{n+1} depend only on q_0 and Γ_0. In this fashion, we recursively obtain an estimate of the form

$$\sum_{i=1}^{k+1} \|q\|_{i,\infty,k-i+1} \leq C\left(\sum_{i=1}^{k+1} \|f\|_{i,1,k-i+1} + \sum_{i=1}^{k} \|g\|_{i,1,k-i} \sum_{i=1}^{k+1} \|q\|_{i,\infty,k-i+1}\right) + C_0, \tag{15}$$

where C_0 depends only on the initial conditions. We can bound the right-hand side of (15) by

$$C\sqrt{T}\left(\sum_{i=1}^{k+1} \|f\|_{i,2,k-i+1} + \sum_{i=1}^{k} \|g\|_{i,2,k-i} \sum_{i=1}^{k+1} \|q\|_{i,\infty,k-i+1}\right) + C_0, \tag{16}$$

and by using this and choosing T sufficiently small, we finally obtain a bound of the form

$$\sum_{i=1}^{k+1} \|q\|_{i,\infty,k-i+1} \leq C\sqrt{T} \sum_{i=1}^{k+1} \|f\|_{i,2,k-i+1} + C_0. \tag{17}$$

Finally, we can get a bound on the $(k+1)$st time derivative of q from the equation itself, and we find that $\|q\|_{0,2,k+1}$ is also bounded by an expression of the form (17).

The second equation of (10) has the form

$$\Gamma_t^{n+1} + (s^n \Gamma_x^{n+1})_x = 0; \tag{18}$$

since we also need to consider time derivatives of this equation, we look at the more general form

$$\Gamma_t + (s\Gamma_x)_x = v, \tag{19}$$

with given s and v. Here s is negative and proportional to $1-x$ as $x \to 1$. We multiply (19) with $(s\Gamma_x)_x$ and integrate. After an integration by parts this yields

$$-\frac{1}{2}\int_0^1 s(x,t)\Gamma_x(x,t)^2\,dx + \frac{1}{2}\int_0^1 s(x,0)\Gamma_0'(0)^2\,dx + \frac{1}{2}\int_0^t\int_0^1 s_t(x,\tau)\Gamma_x(x,\tau)^2\,dx\,d\tau$$

$$+\int_0^t\int_0^1 [(s(x,\tau)\Gamma_x(x,\tau))_x]^2\,dx\,d\tau = \int_0^t\int_0^1 v(x,\tau)(s(x,\tau)\Gamma_x(x,\tau))_x\,dx\,d\tau. \tag{20}$$

Noting that

$$\|\Gamma_x\|_{L^2(0,1)} \le C\|(1-x)\Gamma_x\|_{H^1(0,1)}, \tag{21}$$

we can use (20) to get a bound on $\|\Gamma\|_{2,2,0}$ in terms of v and s:

$$\|\Gamma\|_{2,2,0} \le C\|v\|_{L^2((0,T)\times(0,1))} + C_0. \tag{22}$$

We use this bound on the second equation of (10) and its first k time derivatives. After estimating time derivatives of Γ^{n+1}, we can get bounds for spatial derivatives by exploiting the equation and (21).

The lemma follows by combining the estimates sketched above. We omit the tedious but straightforward details.

On Z_M, we now define a distance function by using a weaker norm than that in (12):

$$d((q,\Gamma),(\tilde{q},\tilde{\Gamma})) = \|q-\tilde{q}\|_{0,2,k} + \sum_{i=1}^{k}\|q-\tilde{q}\|_{i,\infty,k-i} + \|\Gamma-\tilde{\Gamma}\|_{0,2,k}$$

$$+\sum_{i=2}^{k+1}\|\Gamma-\tilde{\Gamma}\|_{i,2,k+1-i}. \tag{23}$$

We now consider two equations of the form (10) and estimate the difference of the solutions. By using the a priori bounds already established in Lemma 2.1 and using similar estimates as above, one can show the following.

Lemma 2.2. *Let the assumptions be as in Lemma* 2.1. *If T is sufficiently small relative to M, then $\mathcal{S}$ is a contraction in Z_M with the metric defined by d.*

Since Z_M with the metric d is a complete metric space, the mapping $\mathcal{S}$ has a unique fixed point, which is the solution we seek.

3. Stationary fronts with vanishing concentration. We now consider the case where $\Gamma = 0$ at $x = 1$, and $h \neq 0$. For this situation, we substitute $\Gamma(x,t) = (1-x)^2\Phi(x,t)$, $p(x,t) = (1-x)^{-1}q(x,t)$. This leads to the equations

$$q_t + \sigma'((1-x)^2\Phi)\left(-\frac{1}{4}\Phi^{-1/2}q^2(1-x)^2\Phi_x^2 + \frac{3}{2}\Phi^{1/2}\Phi_x q^2(1-x) - 2\Phi^{3/2}q^2\right.$$
$$\left.+\frac{1}{2}\Phi^{1/2}\Phi_x(1-x)^2 qq_x - \Phi^{3/2}(1-x)qq_x\right) = 0,$$
$$\Phi_t - 3q\Phi^{3/2}\sigma'((1-x)^2\Phi)((1-x)\Phi_x - 2\Phi)$$
$$+(1-x)[q\Phi^{3/2}\sigma'((1-x)^2\Phi)((1-x)\Phi_x - 2\Phi)]_x = 0. \tag{24}$$

We note that for $x = 1$, we obtain the system of ODEs

$$\dot{q} = 2\Phi^{3/2}q^2\sigma'(0), \qquad \dot{\Phi} = -6q\Phi^{5/2}\sigma'(0). \tag{25}$$

We find from this that $q^3\Phi$ is constant; i.e., q is a multiple of $\Phi^{-1/3}$. Note that $\sigma'(0)$ is negative, and q and Φ are positive quantities. If we prescribe any strictly positive initial data, then (25) will lead to finite time blow up of $\Phi(1)$. Hence solutions with a stationary front can only exist for a finite time, and eventually the front will begin to advance. This behavior is well known for the porous media equation.

We can transform (24) to a nondegenerate parabolic-hyperbolic system on an infinite interval by using the substitution $1 - x = \exp(-y)$ so that y ranges from 0 to ∞. Note that

$$(1-x)\frac{\partial}{\partial x} = \frac{\partial}{\partial y}, \tag{26}$$

so the transformation indeed leads to a nondegenerate parabolic equation. A local time existence result for smooth solutions of the nondegenerate system can then be established along the lines of [6]. The issue remains whether a solution which is "smooth" as a function of y is also smooth in terms of the original variable x. From (26), we see that, for instance, Γ_x is in L^2 if $\Gamma_y \exp(y/2)$ is in L^2. Hence smoothness with respect to x translates into exponential decay with respect to y. Similarly, higher order smoothness means that the solution has an asymptotic expansion in powers of $\exp(-y)$, with sufficiently rapid decay of the remainder. It is easy to adapt the analysis of [6] to exponentially weighted spaces and thus obtain an existence result for solutions which are smooth in x. It should be stressed, however, that solutions will decay exponentially with respect to y only if such exponential decay is already present in the initial data (i.e., q_0 and Γ_0 exponentially approach a constant as $y \to \infty$). No exponential decay in y (and hence no smoothness in x) is created by the differential equation.

4. Advancing fronts with vanishing concentration. We now consider an advancing front at the location $x = \phi(t)$. We first transform the problem to a fixed domain. Since we impose a no-flux condition at $x = 0$, the total amount of surfactant,

$$M_\Gamma = \int_0^{\phi(t)} \Gamma(x,t)\,dx, \tag{27}$$

remains constant. We now set

$$y^2 = M_\Gamma - \int_0^x \Gamma(z,t)\,dz, \tag{28}$$

and we use y as an independent variable in place of x. Note that y ranges over the fixed interval $[0, \sqrt{M_\Gamma}]$, and the front is now at $y = 0$. We find

$$\frac{\partial y}{\partial x} = -\frac{\Gamma}{2y},$$

$$\frac{\partial y}{\partial t} = -\frac{1}{2y}\int_0^x \Gamma_t(z,t)\,dz = \frac{1}{2y}p\Gamma^{3/2}\sigma(\Gamma)_x = -\frac{1}{4y^2}\Gamma^{5/2}p\sigma(\Gamma)_y. \tag{29}$$

Using these relationships, we can evaluate derivatives using the chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}, \qquad \frac{\partial f}{\partial t} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}. \tag{30}$$

In a slight abuse of notation, $\partial/\partial t$ denotes a time derivative for fixed x on the left-hand side, and a time derivative for fixed y on the right-hand side, of the last equation.

By using (30) in (6), we derive the new set of equations

$$p_t - \frac{1}{8y^2}\Gamma^{5/2}\sigma'(\Gamma)pp_y\Gamma_y - \frac{1}{16y^2}\Gamma^{3/2}\sigma'(\Gamma)p^2\Gamma_y^2 = 0,$$

$$\Gamma_t - \frac{1}{4y^2}\Gamma^{5/2}\sigma'(\Gamma)p\Gamma_y^2 + \frac{\Gamma}{2y}\left[\frac{1}{2y}\Gamma^{5/2}\sigma'(\Gamma)p\Gamma_y\right]_y = 0. \tag{31}$$

At an advancing front, we expect Γ to be proportional to $\phi(t) - x$, and hence proportional to y, while h approaches a finite limit, which makes p proportional to $y^{-1/2}$. We therefore substitute $p = y^{-1/2}q$, $\Gamma = y\Phi$, leading to the new equations

$$q_t - \frac{1}{16}\sigma'(y\Phi)\Phi^{5/2}q^2\Phi_y - \frac{1}{8}\sigma'(y\Phi)\Phi^{7/2}qq_y$$

$$-\frac{y}{8}\sigma'(y\Phi)\Phi^{5/2}qq_y\Phi_y - \frac{y}{16}\sigma'(y\Phi)\Phi^{3/2}q^2\Phi_y^2 = 0,$$

$$\Phi_t - \frac{1}{4}\sigma'(y\Phi)\Phi^{7/2}q\Phi_y - \frac{y}{4}\sigma'(y\Phi)\Phi^{5/2}q\Phi_y^2$$

$$+\frac{\Phi}{4}[\sigma'(y\Phi)\Phi^{5/2}q(y\Phi_y + \Phi)]_y = 0. \tag{32}$$

To complete the specification of the problem, we have initial conditions for Φ and q, which can be derived from (3):

$$\Phi(y,0) = \Phi_0(y), \qquad q(y,0) = q_0(y). \tag{33}$$

The no-flux boundary condition assumes the form

$$\Phi(\sqrt{M_\Gamma},t) + \sqrt{M_\Gamma}\Phi_y(\sqrt{M_\Gamma},t) = 0. \tag{34}$$

In addition, we need a boundary condition for the hyperbolic part of the equation at $y = 0$. To obtain such a boundary condition, we need to consider jump relations across the front. Suppose for simplicity that the front propagates into a film of uniform thickness so that $h = h_*$ ahead of the front. From (2), we find the Rankine–Hugoniot condition

$$-\phi'(t)[h] + \frac{1}{2}\sigma'(0)[h^2\Gamma_x] = 0. \tag{35}$$

Here $[\cdot]$ denotes the amount by which a quantity jumps across the front. The speed of the front is the speed with which the surfactant spreads, i.e.,

$$\phi'(t) = \sigma'(0)h\Gamma_x. \tag{36}$$

With h and Γ_x denoting values behind the front, (35) and (36) yield

$$-h\Gamma_x(h - h_*) + \frac{1}{2}h^2\Gamma_x = 0, \tag{37}$$

which simplifies to

$$h = \Phi^{1/2}q = 2h_*. \tag{38}$$

This is the boundary condition which we prescribe at $y = 0$. By evaluating (32) at $y = 0$, we obtain

$$q_t(0) - \frac{1}{16}\sigma'(0)\Phi^{5/2}(0)q^2(0)\Phi_y(0) - \frac{1}{8}\sigma'(0)\Phi^{7/2}q(0)q_y(0) = 0,$$

$$\Phi_t(0) + \frac{7}{8}\sigma'(0)\Phi^{7/2}(0)q(0)\Phi_y(0) + \frac{1}{4}\sigma'(0)\Phi^{9/2}q_y(0) + \frac{1}{4}\sigma''(0)\Phi^{11/2}q(0) = 0. \tag{39}$$

Moreover, (38) implies

$$\frac{1}{2}q\Phi_t + \Phi q_t = 0. \tag{40}$$

By combining (40) and (39), we find the simple condition

$$\Phi_y(0,t) + \frac{1}{3}\frac{\sigma''(0)\Phi^2(0,t)}{\sigma'(0)} = 0. \tag{41}$$

Although (32) is very similar to (7), we cannot use the same iteration. The crucial difference is that q_x in the second equation of (7) appears only in conjunction with a factor $1 - x$, while q_y in the second equation of (32) appears also without a factor y. It is basically for this reason that we need a different iteration and different estimates. To set up an iterative solution for (32), we first apply the operation

$$\frac{\partial}{\partial t} - \frac{1}{8}\sigma'(y\Phi)\Phi^{7/2}q\frac{\partial}{\partial y} - \frac{y}{8}\sigma'(y\Phi)\Phi^{5/2}\Phi_y q\frac{\partial}{\partial y} \tag{42}$$

to the second equation of (32) and then use the first equation of (32) to eliminate terms involving second derivatives of q and terms involving q_t. This yields an equation of the following form:

$$\Phi_{tt} + \Phi_{yt}\left(\frac{1}{2}\sigma'(y\Phi)\Phi^{7/2}q + yQ_1\right) - \Phi_{yy}\left(\frac{1}{16}(\sigma'(y\Phi))^2\Phi^7q^2 + yQ_2\right)$$

$$+\frac{1}{4}\sigma'(y\Phi)\Phi^{7/2}q(y\Phi_y)_{yt}-\frac{1}{32}(\sigma'(y\Phi))^2\Phi^7q^2(y\Phi_y)_{yy}$$

$$-\frac{y}{8}(\sigma'(y\Phi))^2\Phi^6q^2\Phi_y(y\Phi_y)_{yy}+Q_3=0, \tag{43}$$

where the Q_i are of the form

$$Q_i=Q_i(y,\Phi,q,\Phi_t,\Phi_y,q_y). \tag{44}$$

The iterative construction of solutions now proceeds as follows. We first determine a new iterate q^{n+1} from (32):

$$q_t^{n+1}-\frac{1}{16}\sigma'(y\Phi^n)(\Phi^n)^{5/2}(q^n)^2\Phi_y^n-\frac{1}{8}\sigma'(y\Phi^n)(\Phi^n)^{7/2}q^nq_y^{n+1}$$

$$-\frac{y}{8}\sigma'(y\Phi^n)(\Phi^n)^{5/2}q^nq_y^{n+1}\Phi_y^n-\frac{y}{16}\sigma'(y\Phi^n)(\Phi^n)^{3/2}(q^n)^2(\Phi_y^n)^2=0, \tag{45}$$

subject to the prescribed initial condition and the boundary condition

$$(\Phi^n)^{1/2}q^{n+1}=2h^* \tag{46}$$

at $y=0$. Then we determine a new iterate for Φ from (43) in the form

$$\Phi_{tt}^{n+1}+\Phi_{yt}^{n+1}\left(\frac{1}{2}\sigma'(y\Phi^n)(\Phi^n)^{7/2}q^n+yQ_1^n\right)-\Phi_{yy}^{n+1}\left(\frac{1}{16}(\sigma'(y\Phi^n))^2(\Phi^n)^7(q^n)^2+yQ_2^n\right)$$

$$+\frac{1}{4}\sigma'(y\Phi^n)(\Phi^n)^{7/2}q^n(y\Phi_y^{n+1})_{yt}-\frac{1}{32}(\sigma'(y\Phi^n))^2(\Phi^n)^7(q^n)^2(y\Phi_y^{n+1})_{yy}$$

$$-\frac{y}{32}(\sigma'(y\Phi^n))^2(\Phi^n)^6(q^n)^2\Phi_y^n(y\Phi_y^{n+1})_{yy}+Q_3^n=0, \tag{47}$$

where

$$Q_i^n=Q_i(y,\Phi^n,q^n,\Phi_t^n,\Phi_y^n,q_y^{n+1}). \tag{48}$$

To solve (47), we need initial conditions and the boundary conditions (34) and (41); the latter is implemented in the form

$$\Phi_y^{n+1}(0,t)+\frac{1}{3}\frac{\sigma''(0)(\Phi^n)^2(0,t)}{\sigma'(0)}=0. \tag{49}$$

At each step of the iteration, we must therefore first solve (45), which is simply a first-order hyperbolic PDE. Then one has to solve (47). Before stating a formal result, we outline the basic energy estimate which is used to deal with (47). We represent the equation in the schematic form

$$\Phi_{tt}-(a+yQ_1)\Phi_{yt}-(b+yQ_2)\Phi_{yy}-c(y\Phi_y)_{yt}-d(y\Phi_y)_{yy}+Q_3=0. \tag{50}$$

Moreover, we have the boundary conditions

$$\Phi_y(0,t)=f(t),\qquad \Phi(\sqrt{M_\Gamma},t)+\sqrt{M_\Gamma}\Phi_y(\sqrt{M_\Gamma},t)=0. \tag{51}$$

We note that $d(\sqrt{M_\Gamma}, t) = 0$; the energy estimates will make frequent use of this fact. To obtain the basic energy estimate, we multiply (50) by $\Phi_t + d\Phi_y/c$ and integrate. This yields

$$\sum_{i=1}^{11} A_i = 0. \tag{52}$$

In the following listing of the A_i, it is understood that all integrals over y are from 0 to $\sqrt{M_\Gamma}$.

$$A_1 = \int_0^t \int \Phi_{tt}\Phi_t \, dy \, d\tau = \frac{1}{2}\int \Phi_t^2(y,t)\, dy - \frac{1}{2}\int \Phi_t(y,0)^2 \, dy, \tag{53}$$

$$A_2 = -\int_0^t \int (a + yQ_1)\Phi_{yt}\Phi_t \, dy \, d\tau = \frac{1}{2}\int_0^t a(0,\tau)\Phi_t^2(0,\tau)\, d\tau$$

$$-\frac{1}{2}\int_0^t (a(\sqrt{M_\Gamma},\tau) + \sqrt{M_\Gamma}Q_1(\sqrt{M_\Gamma},\tau))\Phi_t^2(\sqrt{M_\Gamma},\tau)\, d\tau$$

$$+\frac{1}{2}\int_0^t \int (a + yQ_1)_y \Phi_t^2 \, dy \, d\tau, \tag{54}$$

$$A_3 = -\int_0^t \int (b + yQ_2)\Phi_{yy}\Phi_t \, dy \, d\tau = \int_0^t \int (b + yQ_2)\Phi_y \Phi_{yt} \, dy \, d\tau$$

$$+\int_0^t \int (b + yQ_2)_y \Phi_y \Phi_t \, dy \, d\tau + \int_0^t b(0,\tau) f(\tau)\Phi_t(0,\tau)\, d\tau$$

$$-\int_0^t (b(\sqrt{M_\Gamma},\tau) + \sqrt{M_\Gamma}Q_2(\sqrt{M_\Gamma},\tau))\Phi_y(\sqrt{M_\Gamma},\tau)\Phi_t(\sqrt{M_\Gamma},\tau)\, d\tau$$

$$= \frac{1}{2}\int (b(y,t) + yQ_2(y,t))\Phi_y^2(y,t)\, dy - \frac{1}{2}\int (b(y,0) + yQ_2(y,0))\Phi_y^2(y,0)\, dy$$

$$-\frac{1}{2}\int_0^t \int (b + yQ_2)_t \Phi_y^2(y,\tau)\, dy \, d\tau + \int_0^t \int (b + yQ_2)_y \Phi_y \Phi_t \, dy \, d\tau$$

$$+\int_0^t \frac{1}{\sqrt{M_\Gamma}}(b(\sqrt{M_\Gamma},\tau) + \sqrt{M_\Gamma}Q_2(\sqrt{M_\Gamma},\tau))\Phi(\sqrt{M_\Gamma},\tau)\Phi_t(\sqrt{M_\Gamma},\tau)\, d\tau$$

$$+\int_0^t b(0,\tau) f(\tau)\Phi_t(0,\tau)\, d\tau, \tag{55}$$

$$A_4 = \int_0^t \int \frac{d}{c}\Phi_{tt}\Phi_y \, dy \, d\tau = \int \frac{d}{c}(y,t)\Phi_t(y,t)\Phi_y(y,t)\, dy - \int \frac{d}{c}(y,0)\Phi_t(y,0)\Phi_y(y,0)\, dy$$

$$-\int_0^t \int \left(\frac{d}{c}\right)_t \Phi_t\Phi_y \, dy \, d\tau - \int_0^t \int \frac{d}{c}\Phi_t\Phi_{yt}\, dy\, d\tau$$

$$= \int \frac{d}{c}(y,t)\Phi_t(y,t)\Phi_y(y,t)\, dy - \int \frac{d}{c}(y,0)\Phi_t(y,0)\Phi_y(y,0)\, dy$$

$$-\int_0^t \int \left(\frac{d}{c}\right)_t \Phi_t\Phi_y \, dy\, d\tau + \frac{1}{2}\int_0^t \frac{d}{c}(0,\tau)\Phi_t^2(0,\tau)\, d\tau + \frac{1}{2}\int_0^t \int \left(\frac{d}{c}\right)_y \Phi_t^2(y,\tau)\, dy\, d\tau, \tag{56}$$

$$A_5 = -\int_0^t \int (a + yQ_1)\frac{d}{c}\Phi_y\Phi_{yt}\, dy\, d\tau = -\frac{1}{2}\int (a+yQ_1)(y,t)\frac{d}{c}(y,t)\Phi_y^2(y,t)\, dy$$

$$+\frac{1}{2}\int (a+yQ_1)(y,0)\frac{d}{c}(y,0)\Phi_y^2(y,0)\, dy + \frac{1}{2}\int_0^t \int \left[(a+yQ_1)\frac{d}{c}\right]_t \Phi_y^2\, dy\, d\tau, \tag{57}$$

$$A_6 = -\int_0^t \int (b + yQ_2)\frac{d}{c}\Phi_y\Phi_{yy}\, dy\, d\tau = \frac{1}{2}\int_0^t \frac{bd}{c}(0,\tau)\Phi_y^2(0,\tau)\, d\tau$$

$$+\frac{1}{2}\int_0^t \int \left[(b+yQ_2)\frac{d}{c}\right]_y \Phi_y^2\, dy\, d\tau, \tag{58}$$

$$A_7 = -\int_0^t \int c(y\Phi_y)_{yt}\Phi_t\, dy\, d\tau = \int_0^t \int cy\Phi_{yt}^2\, dy\, d\tau + \int_0^t \int c_y y\Phi_{yt}\Phi_t\, dy\, d\tau$$

$$+\int_0^t c(\sqrt{M_\Gamma},\tau)\Phi_t^2(\sqrt{M_\Gamma},\tau)\, d\tau, \tag{59}$$

$$A_8 = -\int_0^t \int d(y\Phi_y)_{yy}\Phi_t\, dy\, d\tau = \int_0^t \int dy\Phi_{yy}\Phi_{yt}\, dy\, d\tau$$

$$+\frac{1}{2}\int d(y,t)\Phi_y^2(y,t)\, dy - \frac{1}{2}\int d(y,0)\Phi_y^2(y,0)\, dy + \int_0^t \int d_y(y\Phi_y)_y\Phi_t\, dy\, d\tau$$

$$+\int_0^t d(0,\tau)f(\tau)\Phi_t(0,\tau)\, d\tau - \frac{1}{2}\int_0^t \int d_t\Phi_y^2\, dy\, d\tau, \tag{60}$$

$$A_9 = -\int_0^t \int d(y\Phi_y)_{yt}\Phi_y\, dy\, d\tau = \int_0^t \int dy\Phi_{yt}\Phi_{yy}\, dy\, d\tau$$

$$+\int_0^t \int d_y y \Phi_{yt} \Phi_y \, dy \, d\tau, \tag{61}$$

$$A_{10} = -\int_0^t \int \frac{d^2}{c} (y\Phi_y)_{yy} \Phi_y \, dy \, d\tau = \int_0^t \int \frac{d^2}{c} y \Phi_{yy}^2 \, dy \, d\tau$$

$$+\int_0^t \int \frac{d^2}{c} \Phi_y \Phi_{yy} \, dy \, d\tau + \int_0^t \int (\frac{d^2}{c})_y (y\Phi_y)_y \Phi_y \, dy \, d\tau$$

$$+\int_0^t \frac{d^2(0,\tau)}{c(0,\tau)} f^2(0,\tau) \, d\tau, \tag{62}$$

$$A_{11} = \int_0^t \int Q_3 \left(\Phi_t + \frac{d}{c} \Phi_y \right) dy \, d\tau. \tag{63}$$

For reasons which will become clear later, we split up Q_3 in the form $Q_3 = A + B_y$. We can then transform (63) further as follows:

$$A_{11} = \int_0^t \int A \left(\Phi_t + \frac{d}{c} \Phi_y \right) dy \, d\tau + \int_0^t \int \left(B_t + \frac{d}{c} B_y \right) \Phi_y \, dy \, d\tau$$

$$+\int_0^t B(\sqrt{M_\Gamma}, \tau) \Phi_t(\sqrt{M_\Gamma}, \tau) \, d\tau - \int_0^t B(0,\tau)\Phi_t(0,\tau) \, d\tau$$

$$-\int B(y,t)\Phi_y(y,t) \, dy + \int B(y,0)\Phi_y(y,0) \, dy. \tag{64}$$

From $\sum A_i$, we can extract the following quadratic terms:

$$\int \frac{1}{2} \Phi_t^2(y,t) + \left(\frac{d}{c} + O(y) \right) \Phi_t \Phi_y(y,t) + \left(\frac{b}{2} - \frac{ad}{2c} + \frac{d}{2} + O(y) \right) \Phi_y^2(y,t) \, dy, \tag{65}$$

$$\int_0^t \int cy \Phi_{yt}^2 + 2dy \Phi_{yy} \Phi_{tt} + \frac{d^2}{c} y \Phi_{yy}^2 \, dy \, d\tau$$

$$= \int_0^t \int cy \left(\Phi_{yt} + \frac{d}{c} \Phi_{yy} \right)^2 dy \, d\tau. \tag{66}$$

From (47), we find that, at $y = 0$,

$$\frac{d}{c} = -\frac{1}{8} \sigma'(0) (\Phi^n)^{7/2} q^n, \tag{67}$$

$$\frac{b}{2} - \frac{ad}{2c} + \frac{d}{2} = \frac{1}{64} \sigma'(0)^2 (\Phi^n)^7 (q^n)^2. \tag{68}$$

As a consequence, the quadratic expression in (65) is positive at least for small y. An additional positive contribution to the energy equation is given by

$$\int_0^t \left(\frac{1}{2}a(0,\tau)+\frac{d(0,\tau)}{2c(0,\tau)}\right)\Phi_t^2(0,\tau)\,d\tau. \tag{69}$$

Most of the remaining terms in the energy identity can be estimated in a fairly straightforward fashion. In A_7 through A_{10}, there are terms involving the product of a second and first derivative of Φ; these terms can be integrated by parts in an analogous fashion as we did for A_1 through A_6; in doing so, we obtain terms involving second derivatives of c and d. We also note that

$$\int_p^{\sqrt{M_\Gamma}} \Phi_y^2(y,t)\,dy \tag{70}$$

for $p>0$ can be estimated in terms of

$$\int_{p(0)}^{\sqrt{M_\Gamma}} \Phi_y^2(0,t)\,dy \tag{71}$$

and

$$\sqrt{t}\int_0^t\int_{p(\tau)}^{\sqrt{M_\Gamma}}\left(\Phi_{yt}+\frac{d}{c}\Phi_{yy}\right)^2(y,\tau)\,dy\,d\tau, \tag{72}$$

where $p(\tau)$ is defined by

$$p'(\tau)=d(p(\tau),\tau)/c(p(\tau),\tau), \qquad p(t)=p. \tag{73}$$

Moreover,

$$\int_0^t \Phi_t^2(\sqrt{M_\Gamma},\tau)\,d\tau \tag{74}$$

has a bound of the form

$$\epsilon\int_0^t\int y\left(\Phi_{yt}+\frac{d}{c}\Phi_{yy}\right)^2 dy\,d\tau + C(\epsilon)\int_0^t\int \Phi_t^2+\Phi_y^2\,dy\,d\tau, \tag{75}$$

where ϵ can be chosen arbitrarily small. By using these estimates, one finds, for sufficiently small t, a bound of the form

$$\begin{aligned}&\|\Phi_t\|_{L^\infty((0,t);L^2)}+\|\Phi_y\|_{L^\infty((0,t);L^2)}+\|\sqrt{y}\left(\Phi_{yt}+\frac{d}{c}\Phi_{yy}\right)\|_{L^2((0,t);L^2)}\\ &+\|\Phi_t(0,\cdot)\|_{L^2(0,t)}\le C\left(\|A\|_{L^1((0,t);L^2)}+\|B\|_{L^\infty((0,t),L^2)}+\left\|B_t+\frac{d}{c}B_y\right\|_{L^1((0,t);L^2)}\right)+C_0.\end{aligned} \tag{76}$$

Here C_0 depends only on the initial data, and C depends only on t and on the coefficients and their derivatives up to second order.

We can use the energy estimate to establish the existence of a solution to (50) and (51). For this purpose, we use a Galerkin approximation. Let ψ_m be a basis for $H^1(0, \sqrt{M_\Gamma})$. We then determine an approximate solution Φ_N as follows:

$$(\Phi_N)_t + \frac{d}{c}(\Phi_N)_y = \sum_{m=1}^{N} \alpha_m(t)\psi_m(y), \tag{77}$$

with initial condition

$$\Phi_N(y,0) = \Phi(y,0) \tag{78}$$

and boundary condition

$$(\Phi_N)_y(0,t) = f(t). \tag{79}$$

The α_m are determined from the equations

$$\begin{aligned}
&\int [(\Phi_N)_{tt} - (a+yQ_1)(\Phi_N)_{yt} + (b+yQ_2)_y(\Phi_N)_y + c_y y(\Phi_N)_{yt} - d_{yy}y(\Phi_N)_y]\psi_m\,dy \\
&\quad + \int [(b+yQ_2)(\Phi_N)_y + c(y(\Phi_N)_y)_t + d(y(\Phi_N)_y)_y - d_y y(\Phi_N)_y](\psi_m)_y\,dy \\
&\quad + \psi_m(\sqrt{M_\Gamma})\left[\frac{b+yQ_2}{\sqrt{M_\Gamma}}\Phi_N + c(\Phi_N)_t - d_y\Phi_N\right](\sqrt{M_\Gamma},t) \\
&\quad + \psi_m(0)[b(0,t)f(t) + d(0,t)f(t)] = 0,
\end{aligned} \tag{80}$$

which simply become a set of first-order integrodifferential equations. The initial condition is

$$\sum_{m=1}^{N} \alpha_m(0)\psi_m(y) = \Pi_N\left[\Phi_t(y,0) + \frac{d}{c}\Phi_y(y,0)\right]. \tag{81}$$

Here Π_N is the orthogonal projection in $L^2(0, \sqrt{M_\Gamma})$. We can now repeat the energy estimates above for the discretized system. This leads to a priori estimates which, in the usual fashion, allow us to extract a weakly convergent subsequence from the Φ_N. The limit is the solution we seek.

To obtain higher regularity of solutions to (50), we need to consider time derivatives of the equation. We shall now demonstrate how to obtain a problem of the same form as (50), with Φ_t in place of Φ. Differentiation of (50) with respect to time leads to

$$\begin{aligned}
&\Phi_{ttt} - (a+yQ_1)\Phi_{ytt} - (b+yQ_2)\Phi_{yyt} - c(y\Phi_{yt})_{yt} - d(y\Phi_{yt})_{yy} \\
&\quad -(a+yQ_1)_t\Phi_{yt} - (b+yQ_2)\Phi_{yy} - c_t(y\Phi_y)_{yt} - d_t(y\Phi_y)_{yy} + (Q_3)_t = 0.
\end{aligned} \tag{82}$$

This is indeed of the same form as (50), if we can treat the second line as a new forcing term. The first term presents no problem. To deal with the term involving

Φ_{yy}, we consider (50) as a first-order hyperbolic equation for Φ_{yy} for given Φ_{yt} and Φ_{tt}; this allows us to bound Φ_{yy} in terms of Φ_{yt}, Φ_{tt}, and the initial data. We next use (50) to express $y\Phi_{yyt}$ in terms of $y\Phi_{yyy}$ and second derivatives of Φ; this allows us to eliminate the term $c_t y\Phi_{yyt}$ in (82). Finally, we need to deal with the term $d_t(y\Phi_y)_{yy}$. We write this term as

$$d_t(y\Phi_y)_{yy} = (d_t(y\Phi_y)_y)_y - d_{yt}(y\Phi_y)_y. \tag{83}$$

The second contribution is easily dealt with, and the first is of the form B_y, where (50) can be used to express $B_t + \frac{d}{c}B_y$ in terms of second derivatives of Φ. After the transformations just outlined, (82) is now of the same form as (50) and analogous energy estimates can be applied, with Φ_t in place of Φ.

We are now ready to state an existence result for the initial-boundary value problem consisting of (32)–(34) and (38). To define the function spaces, let $\|\cdot\|_{k,p,l}$ denote the norm in $W^{l,p}((0,T);H^k(0,\sqrt{M_\Gamma}))$. We define Z_M as the set of all (q,Φ) such that

$$q \in \bigcap_{i=0}^{k} W^{k-i,\infty}((0,T);H^i(0,\sqrt{M_\Gamma})),$$

$$\Phi \in \bigcap_{i=0}^{k+1} W^{k+1-i,\infty}((0,T);H^i(0,\sqrt{M_\Gamma})),$$

$$\sum_{i=0}^{k} \|q\|_{i,\infty,k-i} + \sum_{i=0}^{k+1} \|\Phi\|_{i,\infty,k+1-i} \le M,$$

q and its first $k-1$ time derivatives satisfy the appropriate initial conditions,

Φ and its first k time derivatives satisfy the appropriate initial conditions,

$$\Phi \text{ satisfies (34).} \tag{84}$$

By $\mathcal{S}$ we denote the mapping from (q^n,Φ^n) to (q^{n+1},Φ^{n+1}) under the iteration defined above.

The existence of solutions now follows from the following theorem.

THEOREM 4.1. *Let $k \ge 3$, and assume Φ_0 and q_0 are strictly positive, sufficiently smooth, and compatible with the boundary conditions. Moreover, assume that the initial values for time derivatives of Φ and q up to orders $k-1$ and, respectively, k are compatible with the boundary conditions* (38) *and that the initial values for time derivatives of Φ up to order $k-1$ are compatible with* (34) *and* (41). *Finally, assume that the function $\Gamma \to \sigma(\Gamma)$ is smooth and $\sigma' < 0$. If M is chosen sufficiently large and T is sufficiently small relative to M, then $\mathcal{S}$ is a contraction in Z_M under the metric defined by*

$$d((q,\Phi),(\tilde{q},\tilde{\Phi})) = \sum_{i=0}^{k-1} \|q-\tilde{q}\|_{i,\infty,k-1-i} + \sum_{i=0}^{k} \|\Phi-\tilde{\Phi}\|_{i,\infty,k-i}. \tag{85}$$

The proof follows from the energy estimates given above, and we shall only sketch an outline. At each step of the iteration, we must first solve the first-order hyperbolic equation (45). An energy estimate for this equation gives a bound for $\|q^{n+1}\|_{0,\infty,0}$. To get bounds on derivatives of q^{n+1}, we differentiate (45) with respect to t and y and repeat the energy estimate. Note that boundary values for spatial derivatives of q^{n+1} at $y = 0$ can be obtained from the equation (45). We then apply the energy estimates from above to (47) and its time derivatives. In doing so, we take advantage of (64) in dealing with terms involving the highest derivatives of q^{n+1}. This yields bounds for $\|\Phi^{n+1}\|_{0,\infty,1} + \|\Phi^{n+1}\|_{1,\infty,0}$ and analogous bounds for time derivatives of Φ^{n+1}. Once time derivatives have been estimated, it is easy to get bounds on spatial derivatives from (47), which can be regarded as a first-order hyperbolic equation for Φ_{yy}, once Φ_{tt} and Φ_{yt} are known. The result now follows by putting together the bounds obtained in this fashion.

REFERENCES

[1] S. Angenent, *Analyticity of the interface of the porous media equation after the waiting time*, Proc. Amer. Math. Soc., 102 (1988), pp. 329–336.

[2] D. G. Aronson and J. L. Vazquez, *Eventual C^∞-regularity and concavity of flows in one-dimensional porous media*, Arch. Rational Mech. Anal., 99 (1988), pp. 329–348.

[3] M. S. Borgas and J. B. Grotberg, *Monolayer flow on a thin film*, J. Fluid Mech., 193 (1988), pp. 151–170.

[4] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[5] K. Höllig and H.-O. Kreiss, *C^∞-regularity for the porous medium equation*, Math. Z., 192 (1986), pp. 217–224.

[6] M. Renardy, *On an equation describing the spreading of surfactants on thin films*, Nonlinear Anal., 26 (1996), pp. 1207–1219.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1064–1085, September 1997

004

SYMMETRIES AND DECAY OF THE GENERALIZED KADOMTSEV–PETVIASHVILI SOLITARY WAVES*

ANNE DE BOUARD† AND JEAN-CLAUDE SAUT†

Abstract. We prove that the solitary waves of the generalized Kadomtsev–Petviashvili equations, when they exist, are cylindrical with respect to the transverse variables and decay with an optimal algebraic rate.

Key words. Kadomtsev–Petviashvili equations, solitary waves, symmetry, decay

AMS subject classifications. 35B40, 35B99, 35Q53

PII. S0036141096297662

1. Introduction.

This paper is a continuation of a previous work [5] where we proved existence and nonexistence results for solitary waves of the generalized Kadomtsev–Petviashvili (KP) equations

$$(1.1)\qquad \begin{cases} u_t + u^p u_x + u_{xxx} - v_y = 0, \quad u = u(x,y,t),\ (x,y) \in \mathbb{R}^2,\ t > 0, \\ v_x = u_y \end{cases}$$

in the two-dimensional case, and

$$(1.2)\qquad \begin{cases} u_t + u^p u_x + u_{xxx} - v_y - w_z = 0, \quad u = u(x,y,z,t),\ (x,y,z) \in \mathbb{R}^3,\ t > 0, \\ v_x = u_y, \\ w_x = u_z \end{cases}$$

in the three-dimensional case.

In (1.1) (resp., (1.2)), x is the direction of propagation while y (resp., (y,z)) are the transverse variables.

Throughout the paper, we will assume that $p = m/n$, m and n are relatively prime, and n is odd, so that we can define u^p for any real-valued function u.

By solitary wave, we mean (see [5]) a solution of (1.1) (resp., (1.2)) of the type $u(x-ct,y)$ (resp., $u(x-ct,y,z)$), where $c > 0$, and $u \in Y$, with

$$Y = \text{closure of the space } \partial_x(C_0^\infty(\mathbb{R}^d)),\ d = 2,3, \text{ for the norm}$$
$$\|\partial_x\varphi\|_Y = \left(\|\nabla\varphi\|_{L^2}^2 + \left\|\partial_x^2\varphi\right\|_{L^2}^2\right)^{1/2}.$$

We are thus considering "localized" solutions to the systems

$$(1.3)\qquad \begin{cases} -cu_x + u^p u_x + u_{xxx} - v_y = 0, \\ v_x = u_y, \end{cases}$$

$$(1.4)\qquad \begin{cases} -cu_x + u^p u_x + u_{xxx} - v_y - w_z = 0, \\ v_x = u_y, \\ w_x = u_z. \end{cases}$$

[*] Received by the editors January 12, 1996; accepted for publication June 19, 1996.
http://www.siam.org/journals/sima/28-5/29766.html

[†] Analyse Numérique et EDP, Université Paris-Sud, Bâtiment 425, 91405 Orsay cedex, France (Anne.Bouard@math.u-psud.fr, Jean-Claude.Saut@math.u-psud.fr).

Note that we may assume $c = 1$ by the simple scale change $\tilde{u}(x, x') = c^{-1/p} u(\frac{x}{\sqrt{c}}, \frac{x'}{c})$, where $x' = y$ (resp., $x' = (y, z)$). Obviously, (1.3), (1.4) imply that u must solve the equation in $\mathbb{R}^d$, $d = 2, 3$,

$$-\Delta u + \partial_x^4 u + \frac{1}{p+1}(u^{p+1})_{xx} = 0, \tag{1.5}$$

where $\Delta = \partial_x^2 + \partial_y^2$ (resp., $\partial_x^2 + \partial_y^2 + \partial_z^2$).

In order to state the results of the present paper we summarize below (and complete slightly) some relevant facts proved in [5] (see also [14] for an existence proof in the two-dimensional case).

THEOREM 1.1.

(i) *The system* (1.1) *(resp.,* (1.2)*) has a nontrivial solitary wave if and only if* $1 \le p < 4$ *(resp.,* $1 \le p < \frac{4}{3}$*).*

(ii) *Any solitary wave* u *is continuous and tends to zero at infinity. In the case where* p *is an integer,*

$$u \in H^\infty(\mathbb{R}^d) = \bigcap_{k \ge 0} H^k(\mathbb{R}^d), \quad d = 2, 3.$$

Proof. We refer to [5], except for (ii), when $3 < p < 4$ (resp., $1 < p < \frac{4}{3}$), which was not considered there. Note that (ii), when $1 < p < 3$, $p \ne 2$, is contained in the proof of Theorem 4.1 in [5].

First we treat the case $d = 2$, and $3 < p < 4$.

As in the proof of Lemma 4.1 in [5], we get from (1.5)

$$\hat{u}(\xi_1, \xi_2) = \frac{\xi_1^2}{|\xi|^2 + \xi_1^4} \hat{g}(\xi_1, \xi_2), \tag{1.6}$$

where $g = -\frac{1}{p+1} u^{p+1}$ and $\hat{f}$ denotes the Fourier transform of f with respect to (x, y).

We will use a reiteration argument. Since $Y \subset L^6(\mathbb{R}^2)$, $g \in L^{6/(p+1)}$. By Lizorkin's theorem [11], $u, u_y, u_{xx} \in L^{6/(p+1)}$ and from [3, Thm. 10.2] we deduce that $u \in L^{p_1}(\mathbb{R}^2)$, $\frac{p+1}{6} - \frac{1}{p_1} = \frac{2}{3}$, that is, $p_1 = \frac{6}{p-3}$. Assume by induction that $u \in L^{p_n}(\mathbb{R}^2)$ with $p_n \ge p_{n-1} \ge p_0 = 6 > \frac{3p}{2}$ (since $p < 4$). Then, $g \in L^{p_n/(p+1)}(\mathbb{R}^2)$, which yields by Lizorkin's theorem $u, u_y, u_{xx} \in L^{p_n/(p+1)}(\mathbb{R}^2)$ and, again by [3], $u \in L^{p_{n+1}}(\mathbb{R}^2)$, where $\frac{p+1}{p_n} - \frac{1}{p_{n+1}} = \frac{2}{3}$ (or $p_{n+1} = +\infty$ if $\frac{p+1}{p_n} < \frac{2}{3}$). Moreover, $\frac{1}{p_{n+1}} - \frac{1}{p_n} = \frac{p}{p_n} - \frac{2}{3} < 0$, and therefore

$$p_{n+1} > p_n \ge p_0 > \frac{3p}{2}.$$

Finally, either p_n converges to a finite value, or it reaches $+\infty$. The only possible finite limit being $\frac{3p}{2}$, p_n necessarily reaches $+\infty$ in a finite number of steps. Moreover, since for any n, $u, u_y, u_{xx} \in L^{p_n/(p+1)}(\mathbb{R}^2)$, one has $\nabla u \in L^q(\mathbb{R}^2)$, $\forall\ q < +\infty$ and by the Sobolev embedding theorem, $u(x, y) \to 0$ as $x^2 + y^2 \to +\infty$.

The proof for $d = 3$, $1 < p < \frac{4}{3}$ follows the same lines. One starts with $u \in L^{p_0}(\mathbb{R}^3)$, ($p_0 = \frac{10}{3}$), $g \in L^{p_0/(p+1)}(\mathbb{R}^3)$, $u, u_y, u_z, u_{xx} \in L^{p_0/(p+1)}(\mathbb{R}^3)$, which implies by Lizorkin's theorem $u \in L^{p_1}(\mathbb{R}^3)$, $\frac{p+1}{p_0} - \frac{1}{p_1} = \frac{2}{5}$. At the nth stage we obtain $\frac{p+1}{p_n} - \frac{1}{p_{n+1}} = \frac{2}{5}$, and one proves again (using $p < \frac{4}{3}$) that p_n must reach $+\infty$ in a finite number of steps. ∎

Let us now describe the content of the present paper. In the second section we prove that the solitary waves constructed in [5] (see Theorem 1.1 above) are cylindrically symmetric, that is, radially symmetric with respect to the transverse coordinates, up to a translation of the origin. Actually, we give alternate variational characterizations of the solitary waves, and we adapt an argument of Lopes [12] to prove the symmetry.

In section 3, we establish sharp (algebraic) decay rates for *any* solitary wave u of (1.1) (resp., (1.2)), namely, $r^2 u \in L^\infty(\mathbb{R}^2)$ (resp., $r^\delta u \in L^2(\mathbb{R}^3)$, $\forall\ \delta,\ 0 < \delta < \frac{3}{2}$), where $r^2 = x^2 + y^2$ (resp., $x^2 + y^2 + z^2$). The sharpness of this decay rate is put on the fore by the lump solitary wave of the KP I equation ((1.1) with $p = 1$) which decays exactly as $\frac{1}{r^2}$ at infinity, and by the fact that for "general"p's, a nontrivial solitary wave does not belong to $L^1(\mathbb{R}^d)$. Actually, our decay estimates are obtained for any solution in Y of the equation (1.5).

The method is reminiscent of that of Bona and Li [4], though we consider here a "nonsmooth" and multidimensional situation. It consists of a careful analysis of the decay properties of solutions to the convolution equation equivalent to (1.6).

In section 4 we indicate briefly how similar results can be obtained for equations having higher dispersion in x.

Finally, we present in an Appendix a global unique continuation theorem which is a key ingredient in proving the symmetry of the solitary wave (see section 2).

2. Symmetry properties of the ground states. In this section, we prove that any ground state solution of (1.3) (resp., (1.4)) is cylindrically symmetric, in the sense that it has radial symmetry with respect to transverse coordinates, up to a translation of the origin. We recall that by transverse coordinates x', we mean coordinates which are transverse to the direction of propagation, that is, $x' = y$ in the two-dimensional case (1.3) and $x' = (y, z)$ in the three-dimensional case (1.4).

A ground state is a solitary wave which minimizes the action

$$S(u) = E(u) + \frac{c}{2}\int_{\mathbb{R}^d} |u|^2, \quad d = 2 \text{ or } 3$$

among all the nonzero solutions of (1.3) (resp. (1.4)), where c is the velocity of the solitary wave and E is the energy, defined for $u \in Y$ by

$$E(u) = \frac{1}{2}\int_{\mathbb{R}^2} \left(u_x^2 + \left(D_x^{-1} u_y\right)^2\right) dxdy - \frac{1}{(p+1)(p+2)}\int_{\mathbb{R}^2} u^{p+2} dxdy$$

if $d = 2$, and

$$\begin{aligned} E(u) = {} & \frac{1}{2}\int_{\mathbb{R}^3} \left(u_x^2 + \left(D_x^{-1} u_y\right)^2 + \left(D_x^{-1} u_z\right)^2\right) dxdydz \\ & - \frac{1}{(p+1)(p+2)}\int_{\mathbb{R}^3} u^{p+2} dxdydz \ \text{ if } d = 3. \end{aligned}$$

Note (see [5]) that $D_x^{-1} u_y$ is well defined for $u \in Y$, as the unique element $v \in L^2(\mathbb{R}^d)$ such that $v_x = u_y$. An analogous definition holds for $D_x^{-1} u_z$ if $d = 3$.

Note also that both $E(u)$ and $\int_{\mathbb{R}^d} |u|^2$ are conserved quantities for equations (1.1) and (1.2); i.e., if $u(t)$ is a solution of (1.1) or (1.2), then $E(u(t))$ and $\int_{\mathbb{R}^d} |u(t)|^2$ do not depend on t (see, for example, [13]).

In [5], we proved, for p satisfying the conditions of Theorem 1.1, the existence of nontrivial solutions of equation (1.3) (resp., (1.4)), by considering the minimization problem

$$I_\lambda = \inf\left\{\int_{\mathbb{R}^2}\left[u_x^2 + cu^2 + \left(D_x^{-1}u_y\right)^2\right],\ u\in Y,\ \int_{\mathbb{R}^2} u^{p+2} = \lambda\right\}$$

$$\left(\text{resp., } I_\lambda = \inf\left\{\int_{\mathbb{R}^3}\left[u_x^2 + cu^2 + \left(D_x^{-1}u_y\right)^2 + \left(D_x^{-1}u_z\right)^2\right],\ u\in Y,\ \int_{\mathbb{R}^3} u^{p+2} = \lambda\right\}\right).$$

More precisely, with an appropriate choice of λ, if u^* is a minimum of I_λ, then u^* is a solution of the following equations (depending on whether $d=2$ or 3) in $Y'(\mathbb{R}^d)$ (the dual space of Y in the $L^2(\mathbb{R}^d)$ duality):

$$-u_{xx} + cu + D_x^{-2}u_{yy} = \frac{u^{p+1}}{p+1},\ \text{if } d=2, \tag{2.1}$$

$$-u_{xx} + cu + D_x^{-2}u_{yy} + D_x^{-2}u_{zz} = \frac{u^{p+1}}{p+1},\ \text{if } d=3, \tag{2.2}$$

so that u^* satisfies equation (1.3) if $d=2$ and (1.4) if $d=3$, in $\mathcal{D}'(\mathbb{R}^d)$.

The next lemma shows that the solutions u^* obtained in this way are exactly the ground states of equation (1.3) (resp., (1.4)) and also gives two other characterizations of those solutions, which will appear to be useful in proving the symmetry property.

In order to state the lemma, we define

$$K(u) = \frac{1}{2}\int_{\mathbb{R}^2}\left[cu^2 + \left(D_x^{-1}u_y\right)^2\right] - \frac{1}{(p+1)(p+2)}\int_{\mathbb{R}^2} u^{p+2},\ \text{if } d=2 \tag{2.3}$$

and

$$\begin{aligned} K(u) = &\frac{1}{2}\int_{\mathbb{R}^3}\left[cu^2 + \left(D_x^{-1}u_y\right)^2 + \left(D_x^{-1}u_z\right)^2\right] + \frac{1}{6}\int_{\mathbb{R}^3} u_x^2 \\ &- \frac{1}{(p+1)(p+2)}\int_{\mathbb{R}^3} u^{p+2},\ \text{if } d=3. \end{aligned} \tag{2.4}$$

Lemma 2.1. *Let $d=3$; then there is a real positive numberλ^* such that for $u^*\in Y$, the following assertions are equivalent:*

(i) $\int (u^*)^{p+2} = \lambda^*$ *and u^* is a minimum of I_{λ^*},*

(ii) *u^* is a ground state,*

(iii) $K(u^*) = 0$ *and* $\int (u_x^*)^2 = \inf\left\{\int u_x^2,\ u\in Y,\ u\neq 0,\ K(u)=0\right\}$,

(iv) $K(u^*) = 0 = \inf\left\{K(u),\ u\in Y,\ \int u_x^2 = \int (u_x^*)^2\right\}$.

Let $d=2$; then there is a real positive λ^* such that (i)–(iv) *are equivalent modulo a scale change. More precisely,* (i) *and* (ii) *are equivalent and imply* (iii) *and* (iv), *which are also equivalent; conversely, if u^* satisfies* (iii) *or* (iv), *then there is a positive μ such that $u^*(\frac{\cdot}{\mu})$ is a ground state of equation* (1.3).

Proof. To prove Lemma 2.1, we use techniques which are standard for the bound states of nonlinear Schrödinger equations (see, for example, [6]).

Let u be a minimum of I_λ; then there is a positive Lagrange parameter θ_λ such that u satisfies

$$-u_{xx} + cu + D_x^{-2}u_{yy} = \theta_\lambda\frac{u^{p+1}}{p+1},\ \text{if } d=2,\ \text{or}$$

$$-u_{xx} + cu + D_x^{-2} u_{yy} + D_x^{-2} u_{zz} = \theta_\lambda \frac{u^{p+1}}{p+1}, \text{ if } d = 3.$$

Multiplying these equations by u and integrating by parts yields $I_\lambda = \frac{\theta_\lambda}{p+1} \int u^{p+2} = \frac{\lambda \theta_\lambda}{p+1}$ for each positive λ. Since $I_\lambda = \lambda^{2/(p+2)} I_1$, we get $\theta_\lambda = 1$ by choosing $\lambda = \lambda^* = [(p+1)I_1]^{(p+2)/p}$.

Let us now prove that Lemma 2.1 holds with this choice of λ.

(i)$\Longrightarrow$(iii): Assume that u^* satisfies (i). Let $u \in Y$ with $u \neq 0$ and $K(u) = 0$; let $u_\mu = u\left(\frac{\cdot}{\mu}\right)$, with $\mu = \left(\frac{\int (u^*)^{p+2}}{\int u^{p+2}}\right)^{1/d}$ (note that $K(u) = 0$ implies $\int u^{p+2} > 0$ unless $u = 0$), so that $\int u_\mu^{p+2} = \int (u^*)^{p+2}$, and

$$K(u_\mu) = \left(\frac{1}{2} - \frac{1}{d}\right) \mu^{d-2}(1 - \mu^2) \int u_x^2. \tag{2.5}$$

Since u^* is a minimum of I_{λ^*}, we have $K(u^*) = 0$ (see (2.8) and (2.11) in [5] if $d = 2$ and (2.13)–(2.16) in [5] if $d = 3$]) and, on the other hand,

$$\begin{aligned} K(u^*) + \frac{1}{d} \int (u_x^*)^2 + \frac{1}{(p+1)(p+2)} \int (u^*)^{p+2} \\ \leq K(u_\mu) + \frac{1}{d} \int (u_\mu)_x^2 + \frac{1}{(p+1)(p+2)} \int (u_\mu)^{p+2}; \end{aligned}$$

this implies

$$\frac{1}{d} \int (u_x^*)^2 \leq \frac{\mu^{d-2}}{2} \left(1 - \frac{d-2}{d} \mu^2\right) \int u_x^2 \leq \frac{1}{d} \int u_x^2,$$

and (iii) holds.

(iii)$\Longrightarrow$(ii) (modulo a scale change): if u^* satisfies (iii), then there is a Lagrange parameter θ such that u^* solves the Euler–Lagrange equation

$$cu + D_x^{-2} u_{yy} - \frac{u^{p+1}}{p+1} = \theta u_{xx} \text{ if } d = 2, \text{ and}$$

$$cu + D_x^{-2} u_{yy} + D_x^{-2} u_{zz} - \frac{u^{p+1}}{p+1} = \left(\frac{1}{3} + \theta\right) u_{xx} \text{ if } d = 3.$$

It is easily seen, by multiplying these equations by u^*, integrating by parts, and using $K(u) = 0$, that θ is positive. Hence, setting $u_\mu = u^*(\frac{\cdot}{\mu})$, with $\mu = \frac{1}{\sqrt{\theta}}$ if $d = 2$ (resp., $\mu = 1/\sqrt{\frac{1}{3} + \theta}$ if $d = 3$), u_μ satisfies equation (2.1) (resp., (2.2)).

If $d = 3$, then by (2.5), $K(u_\mu) = \frac{1}{6} \mu (1 - \mu^2) \int (u_x^*)^2$, but since any solution of (2.2) satisfies $K(u) = 0$, this implies $\mu = 1$, i.e., $\theta = \frac{2}{3}$.

Now, the identity $S(u) = K(u) + \frac{1}{d} \int u_x^2$ shows that if u is a solution of (1.3) (resp., (1.4)), then

$$S(u) = \frac{1}{d} \int u_x^2 \geq \frac{1}{d} \int (u_x^*)^2 = \frac{1}{d} \int (u_\mu)_x^2 = S(u_\mu);$$

hence u_μ is a ground state.

(ii) $\Longrightarrow$(i): By the computations in [5, section 2], one has, for any solution u of equation (1.3) (resp., (1.4)), $K(u) = 0$ and

$$\int_{\mathbb{R}^2} \left(u_x^2 + cu^2 + D_x^{-1}u_y^2\right) = \left(1 + \frac{2}{p}\right) \int_{\mathbb{R}^2} u_x^2$$

$$\left(\text{resp., } \int_{\mathbb{R}^3} (u_x^2 + cu^2 + (D_x^{-1}u_y)^2 + (D_x^{-1}u_z)^2 = \left(\frac{4}{3p} + \frac{2}{3}\right) \int_{\mathbb{R}^3} u_x^2\right).$$

Hence if u^* is a ground state, u^* minimizes both $\int u_x^2$ and $\int_{\mathbb{R}^2}(u_x^2 + cu^2 + (D_x^{-1}u_y)^2)$ (resp., $\int_{\mathbb{R}^3}(u_x^2 + cu^2 + (D_x^{-1}u_y)^2 + (D_x^{-1}u_z)^2)$) among all the solutions of (1.3) (resp. (1.4)). Let $\lambda = \int_{\mathbb{R}^d} u^{p+2}$ and $\tilde{u}$ be a minimum of I_λ. Then

$$I_\lambda = \int_{\mathbb{R}^d} \left(\tilde{u}_x^2 + c\tilde{u}^2 + (D_x^{-1}\tilde{u}_y)^2\right) \le \int_{\mathbb{R}^d} \left((u_x^*)^2 + c(u^*)^2 + (D_x^{-1}u_y^*)^2\right)$$

and there is a positive θ such that

$$c\tilde{u} + D_x^{-2}\tilde{u}_{yy} - \tilde{u}_{xx} = \theta \frac{\tilde{u}^{p+1}}{p+1}.$$

Using the equations satisfied by $\tilde{u}$ and u^*, the preceding inequality is written as

$$I_\lambda = \frac{\theta\lambda}{p+1} \le \frac{\lambda}{p+1};$$

hence $\theta \le 1$.

On the other hand, $\bar{u} = \theta^p \tilde{u}$ satisfies equation (1.3) (resp., (1.4)), and since u^* is a ground state,

$$\begin{aligned}\int_{\mathbb{R}^d} \left((u_x^*)^2 + c(u^*)^2 + (D_x^{-1}u_y^*)^2\right) &\le \int_{\mathbb{R}^d} \left(\bar{u}_x^2 + c\bar{u}^2 + (D_x^{-1}\bar{u}_y)^2\right) \\ &\le \theta^{2p} \int_{\mathbb{R}^d} \left(\tilde{u}_x^2 + c\tilde{u}^2 + (D_x^{-1}\tilde{u}_y)^2\right)\end{aligned}$$

so that $\theta \ge 1$.

Hence $u^* = \tilde{u}$ is a minimum of I_λ with $\lambda = \lambda^*$.

(iii)$\Longleftrightarrow$(iv): Assume that (iii) holds; let $u \in Y$ with $\int_{\mathbb{R}^d} u_x^2 = \int_{\mathbb{R}^d} (u_x^*)^2$. Note that $K(\eta u) > 0$ for $\eta > 0$ sufficiently small, so that if $K(u) < 0$, then there is an $\eta_0 \in (0, 1)$ such that $K(\eta_0 u) = 0$; then setting $\tilde{u} = \eta_0 u$, one has $\tilde{u} \in Y$, $K(\tilde{u}) = 0$ and $\int_{\mathbb{R}^d} (\tilde{u}_x)^2 < \int_{\mathbb{R}^d} u_x^2 = \int_{\mathbb{R}^d} (u_x^*)^2$, which contradicts (iii), and shows that u^* satisfies (iv) since $K(u^*) = 0$.

On the opposite, assume that u^* satisfies (iv) and let $u \in Y$ with $K(u) = 0$, $u \ne 0$. Then $K(\eta u) < 0$ for $\eta > 1$, so if $\int u_x^2 < \int (u_x^*)^2$, one can find $\eta_0 > 1$ with $\int (\eta_0 u)_x^2 = \int (u_x^*)^2$ and $K(\eta_0 u) < 0$, contradicting (iv). Hence $\int u_x^2 \ge \int (u_x^*)^2$ and (iii) holds. This ends the proof of Lemma 2.1. ☐

We now state and prove our theorem concerning the symmetry properties of the ground state solutions of equation (1.3) (resp., (1.4)).

THEOREM 2.1. *Let $x' = y \in \mathbb{R}$ if $d = 2$ and $x' = (y, z) \in \mathbb{R}^2$ if $d = 3$; then, up to a translation of the origin of coordinates in x', any ground state u^* is radial in x'; that is, u^* only depends on x and $|x'|$.*

Proof of Theorem 2.1. We use an argument of Lopes [12].

Case $d = 2$. Choose $b \in \mathbb{R}$, in order that if $\Delta = \{(x, y) \in \mathbb{R}^2, y = b\}$, then

$$\int_{\Delta^+} (u_x^*)^2 = \int_{\Delta^-} (u_x^*)^2 = \frac{1}{2} \int_{\mathbb{R}^2} (u_x^*)^2 ,$$

where Δ^+ and Δ^- are the half-planes delimited by Δ. Let u^+ be defined by $u^+ = u^*$ in Δ^+ and u^+ be symmetric with respect to Δ. Then $u^+ \in Y$; indeed, if $\varphi \in L^2_{\text{loc}}$ is such that $\varphi_x = u^*$ and $\varphi_y = D_x^{-1} u_y^*$, and if

$$\varphi^+(x, y) = \begin{cases} \varphi(x, y) & \text{if } y > b, \\ \varphi(x, 2b - y) & \text{if } y < b, \end{cases}$$

then $\varphi_x^+ = u^+$ and $\int_{\mathbb{R}^2} \left(\varphi_y^+\right)^2 = 2 \int_{\Delta^+} \varphi_y^2 < +\infty$. Since there is a sequence $\varphi_n \in \mathcal{C}_0^\infty(\mathbb{R}^2)$ such that $(\varphi_n)_x$ converges to $\varphi_x = u^*$ in Y, $D_x^{-1} u_y^+ = \varphi_y^+$. Moreover, $\int_{\mathbb{R}^2} \left(u_x^+\right)^2 = \int_{\mathbb{R}^2} \left(u_x^*\right)^2$. In the same way, if $u^- = u^*$ in Δ^- and u^- is symmetric with respect to Δ, then $u^- \in Y$ and $\int_{\mathbb{R}^2} \left(u_x^-\right)^2 = \int_{\mathbb{R}^2} \left(u_x^*\right)^2$. Hence it follows from Lemma 2.1 (iv) that $K(u^+) \geq 0$ and $K(u^-) \geq 0$.

But one easily checks that

$$K(u^+) + K(u^-) = 2K(u^*) = 0$$

so that u^+ and u^- both satisfy assertion (iv) of Lemma 2.1; it results from this last lemma that u^+ and u^- are ground states of equation (1.3); thus u^+, u^-, and u^* satisfy

$$-u_{xxxx} + u_{xx} + u_{yy} = \left(\frac{u^{p+1}}{p+1} \right)_{xx} \quad \text{in } \mathbb{R}^2.$$

At last, since $u^+ = u^*$ in Δ^+ and $u^- = u^*$ in Δ^-, the unique continuation principle (see the appendix) applied to $u^+ - u^*$ (resp., $u^- - u^*$) tells us that $u^+ = u^- = u^*$, and u^* is symmetric with respect to Δ.

Case $d = 3$. Consider any plane Π parallel to the x axis; then there is a (unique) plane $\tilde{\Pi}$ parallel to Π such that

$$\int_{\tilde{\Pi}^+} (u_x^*)^2 = \int_{\tilde{\Pi}^-} (u_x^*)^2 = \frac{1}{2} \int_{\mathbb{R}^2} (u_x^*)^2 .$$

One can then show, exactly as in dimension 2 (by using the unique continuation principle stated in the appendix) that u^* is symmetric with respect to $\tilde{\Pi}$. It follows that, after a change of origin of transverse coordinates, u^* is symmetric with respect to the coordinate planes containing the x-axis.

It remains to show, using the arguments in [12] or [7], that u^* is symmetric with respect to any plane containing the x-axis. Suppose this is not the case, and let Π be a plane containing the x-axis such that u^* is not symmetric with respect to Π. Let $\tilde{\Pi}$ be parallel to Π such that u^* is symmetric with respect to $\tilde{\Pi}$; then one can construct as in [12] or [7] a sequence (C_n) of cylinders where C_0 is delimited by $\tilde{\Pi}$ and the planes of coordinates containing the x-axis, and the other cylinders are obtained by successive reflexions with respect to the three planes delimiting C_0. For each n, one has $\int_{C_n} (u^*)^2 = \int_{C_0} (u^*)^2$, and there is a subsequence of (C_n) consisting of disjoint cylinders; since u^* is square integrable, this implies $u^* = 0$ on C_0; hence, $u^* = 0$ on each C_n, but, then, $u^* = 0$ everywhere, since $\cup_{n \in \mathbb{N}} C_n = \mathbb{R}^3$.

This ends the proof of Theorem 2.1. □

3. Algebraic decay of the solitary waves. To start, we give an optimal result for the decay of the solitary wave in the two-dimensional case.

THEOREM 3.1. *Any nontrivial solitary wave of* (1.1) *satisfies*

$$r^2 u \in L^\infty(\mathbb{R}^2), \quad r^2 = x^2 + y^2. \tag{3.1}$$

Remark 3.1. Theorem 3.1 is sharp in two ways. First, the lump solution of the KP I equation ((1.1) with $p = 1$), namely,

$$u(x-ct,y) = \frac{8c\left(1 - \frac{c}{3}(x-ct)^2 + \frac{c^2}{3}y^2\right)}{\left(1 + \frac{c}{3}(x-ct)^2 + \frac{c^2}{3}y^2\right)^2}, \tag{3.2}$$

shows that in the two-dimensional case one cannot expect a decay rate better than r^{-2}. On the other hand (and this is valid for $d = 2, 3$), writing

$$-(p+1)\hat{u}(\xi_1, \xi_\perp) = \frac{\xi_1^2}{|\xi|^2 + \xi_1^4}\widehat{u^{p+1}}(\xi_1, \xi_\perp), \quad \xi_\perp = \xi_2 \text{ (resp., } (\xi_2, \xi_3)) \tag{3.3}$$

shows that u cannot belong to $L^1(\mathbb{R}^d)$, when $p = \frac{m}{n}$, m odd. (Since $\xi \mapsto \frac{\xi_1^2}{|\xi|^2+\xi_1^4}$ is not continuous at the origin, this would lead to the absurd conclusion that $\int_{\mathbb{R}^d} u^{p+1} = 0$.) So, in this case, if the solitary wave decays with an algebraic rate $r^{-\alpha}$, then necessarily $\alpha \leq d$.

Proof of Theorem 3.1. To prove (3.1), we start with a simple integral decay estimate and then use the convolution equation equivalent to (1.6).

LEMMA 3.1. *Any solitary wave of* (1.1) *satisfies*

$$\int_{\mathbb{R}^2} \left(x^2 + y^2\right)\left(|\nabla u|^2 + u_{xx}^2\right) dxdy < +\infty. \tag{3.4}$$

Proof of Lemma 3.1. Let $\chi_0 \in \mathcal{C}_0^\infty(\mathbb{R})$, $0 \leq \chi_0 \leq 1$, $\chi_0(t) = 1$ if $0 \leq |t| \leq 1$, $\chi_0(t) = 0$, $|t| \geq 2$. We set $\chi_j(x) = \chi_0\left(\frac{x^2}{j^2}\right)$, $j = 1, 2, \ldots.$ We multiply (1.5) by $\chi_j(x)x^2u$ and integrate over $\mathbb{R}^2$. Using several integrations by parts, the terms in (1.5) are computed as follows.

$$\begin{aligned}
-\int_{\mathbb{R}^2} u_{xx}\chi_j(x)x^2u &= 2\int_{\mathbb{R}^2} x\chi_j(x)uu_x + \int_{\mathbb{R}^2} \chi_j'(x)x^2uu_x + \int_{\mathbb{R}^2} x^2\chi_j(x)u_x^2 \\
&= \int_{\mathbb{R}^2} x^2\chi_j(x)u_x^2 - \int_{\mathbb{R}^2} \chi_j(x)u^2 - \int_{\mathbb{R}^2} \left(2x\chi_j'(x) + \frac{1}{2}\chi_j''(x)\right)u^2,
\end{aligned}$$

$$-\int_{\mathbb{R}^2} u_{yy}x^2\chi_j(x)u = \int_{\mathbb{R}^2} x^2\chi_j(x)u_y^2,$$

$$\begin{aligned}\int_{\mathbb{R}^2} x^2\chi_j(x)uu_{xxxx} &= \int_{\mathbb{R}^2}(x^2u\chi_j(x))_{xx}u_{xx}\\ &= \int_{\mathbb{R}^2}(2\chi_j(x)+4x\chi_j'(x)+x^2\chi_j''(x))uu_{xx}\\ &+ \int_{\mathbb{R}^2}(4x\chi_j(x)+2x^2\chi_j'(x))u_xu_{xx}+\int_{\mathbb{R}^2} x^2\chi_j(x)u_{xx}^2\\ &= \int_{\mathbb{R}^2} x^2\chi_j(x)u_{xx}^2-4\int_{\mathbb{R}^2}\chi_j(x)u_x^2-2\int_{\mathbb{R}^2}(4x\chi_j'(x)+x^2\chi_j''(x))u_x^2\\ &+\int_{\mathbb{R}^2}\left(6\chi_j''(x)+4x\chi_j'''(x)+\frac{1}{2}x^2\chi_j^{(4)}(x)\right)u^2,\end{aligned}$$

$$\begin{aligned}\int_{\mathbb{R}^2} x^2\chi_j(x)u\left(u^{p+1}\right)_{xx} &= 2\frac{p+1}{p+2}\int_{\mathbb{R}^2}\chi_j(x)u^{p+2}-(p+1)\int_{\mathbb{R}^2}x^2\chi_j(x)u^pu_x^2\\ &+\frac{p+1}{p+2}\int_{\mathbb{R}^2}(x^2\chi_j''(x)+4x\chi_j'(x))u^{p+2}.\end{aligned}$$

Finally, we arrive at

$$\begin{aligned}\int_{\mathbb{R}^2} x^2\chi_j(x)\left(|\nabla u|^2+u_{xx}^2\right) &= \int_{\mathbb{R}^2}\chi_j(x)\left[u^2+4u_x^2-\frac{2}{p+2}u^{p+2}\right]\\ &+\int_{\mathbb{R}^2}x^2\chi_j(x)u_x^2u^p+2\int_{\mathbb{R}^2}(4x\chi_j'(x)+x^2\chi_j''(x))u_x^2\\ &-\frac{1}{p+2}\int_{\mathbb{R}^2}(4x\chi_j'(x)+x^2\chi_j''(x))u^{p+2}\\ &+\int_{\mathbb{R}^2}\left(2x\chi_j'(x)+\frac{1}{2}x^2\chi_j''(x)-6\chi_j''(x)-4x\chi_j'''(x)-\frac{1}{2}x^2\chi_j^{(4)}(x)\right)u^2.\end{aligned} \tag{3.5}$$

In a similar fashion, we multiply (1.5) by $\chi_j(y)y^2u$ and integrate over $\mathbb{R}^2$ to get, successively,

$$-\int_{\mathbb{R}^2} y^2u_{xx}\chi_j(y)u=\int_{\mathbb{R}^2}y^2\chi_j(y)u_x^2,$$

$$\begin{aligned}-\int_{\mathbb{R}^2} y^2\chi_j(y)u_{yy}u &= \int_{\mathbb{R}^2}y^2\chi_j(y)u_y^2-\int_{\mathbb{R}^2}\chi_j(y)u^2\\ &-\int_{\mathbb{R}^2}\left(2y\chi_j'(y)+\frac{1}{2}y^2\chi_j''(y)\right)u^2,\end{aligned}$$

$$\int_{\mathbb{R}^2} y^2\chi_j(y)u_{xxxx}u=\int_{\mathbb{R}^2}y^2\chi_j(y)u_{xx}^2,$$

$$\frac{1}{p+1}\int_{\mathbb{R}^2}\chi_j(y)\left(u^{p+1}\right)_{xx}y^2u=-\int_{\mathbb{R}^2}u_x^2\chi_j(y)y^2u^p.$$

Adding these inequalities yields

$$\begin{aligned}\int_{\mathbb{R}^2} y^2\chi_j(y)\left(|\nabla u|^2+u_{xx}^2\right) &= \int_{\mathbb{R}^2}\chi_j(y)u^2+\int_{\mathbb{R}^2} y^2\chi_j(y)u^p u_x^2\\ &\quad+\int_{\mathbb{R}^2}\left(2y\chi_j'(y)+\frac{1}{2}y^2\chi_j''(y)\right)u^2.\end{aligned}\tag{3.6}$$

Finally, (3.5) and (3.6) imply

$$\begin{aligned}&\int_{\mathbb{R}^2}\left[x^2\chi_j(x)+y^2\chi_j(y)\right]\left(|\nabla u|^2+u_{xx}^2\right)\\ &=4\int_{\mathbb{R}^2}\chi_j(x)u_x^2+\int_{\mathbb{R}^2}\left[\chi_j(x)+\chi_j(y)\right]u^2-\frac{2}{p+2}\int_{\mathbb{R}^2}\chi_j(x)u^{p+2}\\ &+\int_{\mathbb{R}^2}\left[x^2\chi_j(x)+y^2\chi_j(y)\right]u_x^2u^p\\ &+\int_{\mathbb{R}^2}\left[2x\chi_j'(x)+2y\chi_j'(y)+\frac{1}{2}x^2\chi_j''(x)+\frac{1}{2}y^2\chi_j''(y)\right]u^2\\ &-\int_{\mathbb{R}^2}\left[6\chi_j''(x)+4x\chi_j'''(x)+\frac{1}{2}x^2\chi_j^{(4)}(x)\right]u^2\\ &+2\int_{\mathbb{R}^2}\left[4x\chi_j'(x)+x^2\chi_j''(x)\right]u_x^2-\frac{1}{p+2}\int_{\mathbb{R}^2}\left[4x\chi_j'(x)+x^2\chi_j''(x)\right]u^{p+2}.\end{aligned}\tag{3.7}$$

Let us consider the right-hand side of (3.7). The first three terms tend, as $j\to+\infty$, to $2\int_{\mathbb{R}^2}\left[2u_x^2+u^2-\frac{1}{p+2}u^{p+2}\right]$ by Lebesgue's theorem. The terms involving derivatives of χ_j tend to zero, again by Lebesgue's theorem and the properties of χ_j.

On the other hand, since $u\to0$ as $r\to+\infty$, there exists $R>0$ such that $r\geq R$ implies $|u^p|\leq\frac{1}{2}$. Thus

$$\int_{\mathbb{R}^2}\left[x^2\chi_j(x)+y^2\chi_j(y)\right]u_x^2u^p\leq C(R)+\frac{1}{2}\int_{\mathbb{R}^2}\left[x^2\chi_j(x)+y^2\chi_j(y)\right]u_x^2,$$

and finally, (3.7) implies that

$$\int_{\mathbb{R}^2}\left[x^2\chi_j(x)+y^2\chi_j(y)\right]\left(|\nabla u|^2+u_{xx}^2\right)$$

is uniformly bounded in j. Our claim follows from Fatou's lemma. □

As we mentioned previously, our analysis of the decay of the solitary wave is based on the convolution equation

$$u=ih*(u^pu_x),\tag{3.8}$$

where

$$\hat{h}(\xi_1,\xi_2)=\frac{\xi_1}{|\xi|^2+\xi_1^4},\quad |\xi|^2=\xi_1^2+\xi_2^2.$$

Lemma 3.2. *There exists a constant $C>0$ such that*

$$|h(x,y)|\leq\frac{C}{r},\quad\forall\ (x,y)\in\mathbb{R}^2\ \text{ where } r=(x^2+y^2)^{1/2}.$$

Proof. We have

$$h(x,y)=\int_{\mathbb{R}^2}\frac{\xi_1}{\xi_1^2+\xi_2^2+\xi_1^4}e^{ix\xi_1+iy\xi_2}d\xi_1 d\xi_2;$$

writing

$$\frac{\xi_1}{\xi_1^2+\xi_2^2+\xi_1^4}=\frac{1}{\xi_1^2\left(1+\xi_1^2\right)\left(\frac{\xi_2^2}{a^2}+1\right)},\quad a^2=\xi_1^2+\xi_1^4,$$

the integral is transformed under the change of variable $\xi_2=a\xi_2'$ into

$$\int_{\mathbb{R}^2}\frac{\text{sgn }\xi_1}{\sqrt{1+\xi_1^2}}\left[\int_{\mathbb{R}^2}\frac{e^{iy|\xi_1|\left(1+\xi_1^2\right)^{1/2}\xi_2'}}{1+\xi_2'^2}d\xi_2'\right]e^{ix\xi_1}d\xi_1;$$

hence, since $\mathcal{F}\left(\frac{1}{1+\xi^2}\right)(y)=e^{-|y|}$,

$$h(x,y)=\int_{-\infty}^{\infty}\frac{\text{sgn }\xi}{\left(1+\xi^2\right)^{1/2}}e^{-|y||\xi|\left(1+\xi^2\right)^{1/2}}e^{ix\xi}\ d\xi.$$

Let us consider first the case $y\neq 0$. Let

$$\begin{aligned}h_1(x,y)&=\int_0^{\infty}\frac{1}{\left(1+\xi^2\right)^{1/2}}e^{-|y|\xi\left(1+\xi^2\right)^{1/2}}e^{ix\xi}d\xi\\&=\int_0^{\infty}\frac{1}{\left(1+\xi^2\right)^{1/2}K'(\xi)}\frac{d}{d\xi}\left[e^{K(\xi)}\right]d\xi,\end{aligned}$$

with $K(\xi)=ix\xi-|y|\xi\left(1+\xi^2\right)^{1/2}$. Thus

$$\begin{aligned}h_1(x,y)&=\left[\frac{e^{K(\xi)}}{ix\left(1+\xi^2\right)^{1/2}-|y|\left(1+2\xi^2\right)}\right]_0^{\infty}-\int_0^{\infty}\frac{d}{d\xi}\left[\frac{1}{\left(1+\xi^2\right)^{1/2}K'(\xi)}\right]e^{K(\xi)}d\xi\\&=\frac{1}{|y|-ix}-\int_0^{\infty}\frac{ix\xi-4|y|\xi\left(1+\xi^2\right)^{1/2}}{\left(1+\xi^2\right)^{1/2}\left[ix\left(1+\xi^2\right)^{1/2}-|y|\left(1+2\xi^2\right)\right]^2}e^{K(\xi)}d\xi\\&\equiv\frac{1}{|y|-ix}-\int_0^{\infty}F(\xi)e^{K(\xi)}d\xi.\end{aligned}$$

Now,

$$\begin{aligned}|F(\xi)|^2&=\frac{x^2\xi^2+16y^2\xi^2\left(1+\xi^2\right)}{\left(1+\xi^2\right)\left[x^2\left(1+\xi^2\right)+y^2\left(1+2\xi^2\right)^2\right]^2}\\&\leq\frac{16}{\left(1+\xi^2\right)\left[x^2\left(1+\xi^2\right)+y^2\left(1+2\xi^2\right)^2\right]}\\&\leq\frac{16}{\left(1+\xi^2\right)^2\left(x^2+y^2\right)},\end{aligned}$$

and (since $e^{K(\xi)} \leq 1$)

$$\left| \int_0^\infty F(\xi) e^{K(\xi)} d\xi \right| \leq \frac{4}{r} \int_0^\infty \frac{d\xi}{1+\xi^2} \leq \frac{C}{r}.$$

Finally,

$$|h_1(x,y)| \leq \frac{1}{||y| - ix|} + \frac{C}{r} = \frac{C+1}{r}, \quad \text{if } y \neq 0.$$

In a similar fashion, one proves that

$$h_2(x,y) = -\int_{-\infty}^0 \frac{1}{\left(1+\xi^2\right)^{1/2}} e^{|y|\xi\left(1+\xi^2\right)^{1/2}} e^{ix\xi} d\xi$$

satisfies

$$|h_2(x,y)| \leq \frac{1}{||y| + ix|} + \frac{C}{r} = \frac{C+1}{r} \quad \text{if } y \neq 0.$$

It remains to consider the case where $y = 0$:

$$h(x,0) = \int_{-\infty}^\infty \frac{\operatorname{sgn} \xi}{\left(1+\xi^2\right)^{1/2}} e^{ix\xi} d\xi = \mathcal{F}\left(\frac{\operatorname{sgn} \xi}{\left(1+\xi^2\right)^{1/2}} \right)(x).$$

Let us check that $xh(x,0)$ is a bounded function. Actually,

$$\frac{d}{d\xi}\left(\frac{\operatorname{sgn} \xi}{\left(1+\xi^2\right)^{1/2}} \right) = \frac{1}{\left(1+\xi^2\right)^{1/2}} \delta + g = \delta + g,$$

where δ is the Dirac mass and $g \in L^1(\mathbb{R})$; thus $xh(x,0) = 1 + \mathcal{F}_x^{-1} g \in L^\infty(\mathbb{R})$. □

We prove now a (nonoptimal) pointwise decay estimate on u.

LEMMA 3.3. $ru \in L^\infty(\mathbb{R}^2)$.

Proof. From (3.8) we obtain

$$\begin{aligned} |r(x,y)u(x,y)| \leq{}& C \int_{\mathbb{R}^2} |h(x-x',y-y')r(x-x',y-y')|\, |u^p u_x(x',y')|\, dx'dy' \\ &+ C \int_{\mathbb{R}^2} |h(x-x',y-y')|\, |r(x',y')u^p u_x(x',y')|\, dx'dy'. \end{aligned} \tag{3.9}$$

By Young's inequality (using Lemma 3.2 and $u^p u_x \in L^1(\mathbb{R}^2)$), the first term in the right-hand side of (3.9) belongs to $L^\infty(\mathbb{R}^2)$. On the other hand, as it is easily checked, $\hat{h} \in L^q(\mathbb{R}^2)$, $1 < q < 2$, so that $h \in L^s(\mathbb{R}^2)$, $2 < s < \infty$. Applying Young's inequality, the second term in the right-hand side of (3.9) is bounded by $C \left\| ru_x \right\|_{L^2} \|h\|_{L^s} \|u\|^p_{L^{2ps/(s-2)}}$ for any s, $2 < s < \infty$, and the result follows from Lemma 3.1. □

We continue the proof of Theorem 3.1. We write

$$u = -\frac{1}{p+1} k * u^{p+1} \text{ where } \hat{k}(\xi_1,\xi_2) = \frac{\xi_1^2}{|\xi|^2 + \xi_1^4}. \tag{3.10}$$

LEMMA 3.4. $\hat{k} \in H^s(\mathbb{R}^2)$, *for any* s, $0 \leq s < 1$.

Proof. First we prove that $\hat{k} \in L^2(\mathbb{R}^2)$. Actually,

$$\int_{\mathbb{R}^2} \frac{\xi_1^4}{\left(|\xi|^2+\xi_1^4\right)^2} d\xi_1 d\xi_2 = \int_{\mathbb{R}} \frac{1}{\left(1+\xi_1^2\right)^2} \left[\int_{\mathbb{R}} \frac{d\xi_2}{\left[1+\frac{\xi_2^2}{\xi_1^2(1+\xi_1^2)}\right]^2} \right] d\xi_1$$

$$= \int_{\mathbb{R}} \frac{|\xi_1|}{\left(1+\xi_1^2\right)^{3/2}} d\xi_1 \int_{\mathbb{R}} \frac{d\xi_2}{\left(1+\xi_2^2\right)^2} < +\infty.$$

On the other hand, one easily checks that $|\nabla \hat{k}(\xi_1, \xi_2)| \le C|\hat{h}(\xi_1, \xi_2)|$, and we already noticed that $\hat{h} \in L^q(\mathbb{R}^2)$, for any q such that $1 < q < 2$. Thus $\partial_{\xi_1}\hat{k}, \partial_{\xi_2}\hat{k} \in L^q(\mathbb{R}^2)$, $1 < q < 2$; that is, $\hat{k}$ belongs to the homogeneous Sobolev space $\dot{H}^1_q(\mathbb{R}^2)$. By Bergh and Löfstrom ([2] Thm. 6.5.1), $\dot{H}^1_q(\mathbb{R}^2) \subset \dot{H}^s_2(\mathbb{R}^2)$ for $s = 2(1-\frac{1}{q})$; i.e., $\hat{k} \in \dot{H}^2_s$ for any $s \in [0, 1)$. Finally, $\hat{k} \in H^s(\mathbb{R}^2)$, $0 \le s < 1$. □

Now we prove an (optimal) integral decay estimate on u.

Lemma 3.5. *For any* δ, $0 \le \delta < 1$, one has

$$r^\delta u \in L^2(\mathbb{R}^2), \tag{3.11}$$

$$r^{1+\delta}\nabla u, \quad r^{1+\delta}u_{xx} \in L^2(\mathbb{R}^2). \tag{3.12}$$

Proof of Lemma 3.5. Coming back to (3.10), we estimate for $\delta > 0$

$$\left|\left(1+r^2\right)^{\delta/2} u\right| \le C \left|\left[\left(1+r^2\right)^{\delta/2} k\right] * u^{p+1}\right| + C \left|\left[\left(1+r^2\right)^{\delta/2} u^{p+1}\right] * k\right|. \tag{3.13}$$

Thanks to Lemma 3.4, we have for any $0 \le \delta < 1$,

$$\left\|\left(1+r^2\right)^{\delta/2} k * u^{p+1}\right\|_{L^2} \le C \left\|\hat{k}\right\|_{H^\delta} \|u\|^{p+1}_{L^{p+1}} \le C.$$

Observe now that $k \in L^q(\mathbb{R}^2)$, $\forall\ q,\ 1 < q \le 2$. In fact,

$$\|k\|_{L^q} \le \left\|\left(1+r^2\right)^{s/2} k\right\|_{L^2} \left\|\frac{1}{\left(1+r^2\right)^{s/2}}\right\|_{L^\alpha}, \quad \frac{1}{q} = \frac{1}{2} + \frac{1}{\alpha}, \tag{3.14}$$

so $q \in (1, 2] \iff \alpha \in (2, +\infty]$.

For a given $q \in (1, 2]$ one can choose $s \in [0, 1)$ such that $s\alpha > 2$ and (3.14) implies our claims. Thus

$$\left\|\left(1+r^2\right)^{\delta/2} u^{p+1} * k\right\|_{L^2} \le C \left\|\left(1+r^2\right)^{\delta/2} u^{p+1}\right\|_{L^\beta} \|k\|_{L^q},$$
$$\frac{1}{q} + \frac{1}{\beta} = \frac{3}{2}, \quad q \in (1, 2), \quad \beta \in (1, 2).$$

But $\left(1+r^2\right)^{\delta/2} u^{p+1} \in L^\beta(\mathbb{R}^2)$ if and only if

$$r^{\delta/(p+1)} u \in L^{\beta(p+1)}(\mathbb{R}^2). \tag{3.15}$$

On the other hand,

$$\int_{\mathbb{R}^2} r^\alpha |u|^\gamma \le \left\|r^\alpha u^{\gamma-2}\right\|_{L^\infty} \|u\|^2_{L^2} \le \left\|r^{\frac{\alpha}{\gamma-2}} u\right\|^{\gamma-2}_{L^\infty} \|u\|^2_{L^2}$$
$$\le C\|u\|^2_{L^2},$$

provided $\alpha \leq \gamma - 2$, by Lemma 3.3. In particular, $r^{(\gamma-2)/\gamma} u \in L^\gamma(\mathbb{R}^2)$, $2 \leq \gamma \leq +\infty$. Let $\gamma = \beta(p+1)$. For any $\delta \in (0,1)$ one can choose $\beta \in (1,2)$ such that

$$\frac{\delta}{p+1} \leq \frac{\gamma - 2}{\gamma}, \quad \text{i.e., } \delta \leq p + 1 - \frac{2}{\beta}, \text{ since } p \geq 1.$$

This proves that $r^\delta u \in L^2(\mathbb{R}^2)$, $0 < \delta < 1$.

Let us check briefly that $r^{1+\delta} \nabla u$, $r^{1+\delta} u_{xx} \in L^2(\mathbb{R}^2)$ for any δ, $0 < \delta < 1$. The proof is very similar to that of Lemma 3.1 and we shall sketch it formally, without the truncation argument. We multiply successively (1.5) by $|x|^{2(1+\delta)} u$, $|y|^{2(1+\delta)} u$, to get after several integrations by parts

$$\begin{aligned} &\int_{\mathbb{R}^2} \left(|x|^{2+2\delta} + |y|^{2+2\delta}\right) \left(|\nabla u|^2 + u_{xx}^2\right) \\ &\quad = (1+2\delta)(2+2\delta) \int_{\mathbb{R}^2} |x|^{2\delta} u_x^2 - (1+2\delta)(2+2\delta) \int_{\mathbb{R}^2} |x|^{2\delta} u u_{xx} \\ &\quad + (1+2\delta)(1+\delta) \int_{\mathbb{R}^2} (|x|^{2\delta} + |y|^{2\delta}) u^2 + \int_{\mathbb{R}^2} \left(|x|^{2+2\delta} + |y|^{2+2\delta}\right) u^p u_x^2 \\ &\quad - \frac{1}{p+2}(1+2\delta)(2+2\delta) \int_{\mathbb{R}^2} |x|^{2\delta} |u|^{p+2}. \end{aligned} \tag{3.16}$$

Consider the right-hand side in (3.16). The first three integrals are bounded by Lemma 3.1 and the fact that $r^\delta u \in L^2(\mathbb{R}^2)$. The last one is bounded by $C\|u\|_{L^\infty}^p \left\|r^\delta u\right\|_{L^2}^2$. Finally,

$$\begin{aligned} &\int \left(|x|^{2+2\delta} + |y|^{2+2\delta}\right) \left|u^p u_x^2\right| \\ &\quad \leq C + \frac{1}{2} \int \left(|x|^{2+2\delta} + |y|^{2+2\delta}\right) u_x^2 \end{aligned}$$

for $r \geq R$, R large enough since $u \to 0$ as $r \to +\infty$. □

We now need a decay estimate on k (see (3.10)).

Lemma 3.6.

$$r^2 k \in L^\infty(\mathbb{R}^2). \tag{3.17}$$

Proof of Lemma 3.6. By definition,

$$k(x,y) = \int_{\mathbb{R}^2} \frac{\xi_1^2}{\xi_1^2 + \xi_2^2 + \xi_1^4} e^{ix\xi_1 + iy\xi_2} d\xi_1 d\xi_2.$$

As in the proof of Lemma 3.2, we find

$$k(x,y) = \int_{-\infty}^{\infty} \frac{|\xi|}{\left(1+\xi^2\right)^{1/2}} e^{-|y||\xi|\left(1+\xi^2\right)^{1/2} e^{ix\xi}} d\xi.$$

Consider first the case where $y \neq 0$ and $\xi \geq 0$:

$$k_1(x,y) = \int_0^\infty \frac{\xi}{\left(1+\xi^2\right)^{1/2}} \frac{1}{K'(\xi)} \frac{d}{d\xi} \left[e^{K(\xi)}\right] d\xi,$$

$$K(\xi) = ix\xi - |y|\xi \left(1+\xi^2\right)^{1/2}.$$

By integration by parts,

$$\begin{aligned} k_1(x,y) &= -\int_0^\infty \frac{d}{d\xi}\left[\frac{\xi}{(1+\xi^2)^{1/2}\,K'(\xi)}\right] e^{K(\xi)} d\xi \\ &= -\int_0^\infty \frac{e^{K(\xi)}}{ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)} d\xi \\ &+ \int_0^\infty \frac{\xi^2\left[ix - 4|y|\left(1+\xi^2\right)^{1/2}\right] e^{K(\xi)}}{(1+\xi^2)^{1/2}\left[ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)\right]^2} d\xi \\ &= I_1 + I_2. \end{aligned}$$

We integrate I_1 by parts to get

$$I_1 = -\left[\frac{e^{K(\xi)}\left(1+\xi^2\right)^{1/2}}{\left[ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)\right]^2}\right]_0^\infty + \int_0^\infty \frac{d}{d\xi} H_1(\xi) e^{K(\xi)} d\xi$$

where

$$H_1(\xi) = \frac{\left(1+\xi^2\right)^{1/2}}{\left[ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)\right]^2},$$

$$I_1 = \frac{1}{(ix-|y|)^2} + \int_0^\infty \frac{d}{d\xi} H_1(\xi) e^{K(\xi)} d\xi. \tag{3.18}$$

One finds that

$$\begin{aligned} \frac{d}{d\xi} H_1(\xi) &= \frac{\xi}{(1+\xi^2)^{1/2}} \frac{1}{\left[ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)\right]^2} \\ &- 2\frac{ix\xi - 4|y|\xi\left(1+\xi^2\right)^{1/2}}{\left[ix\,(1+\xi^2)^{1/2} - |y|\,(1+2\xi^2)\right]^3} = F_1 + F_2. \end{aligned}$$

Obviously,

$$\begin{aligned} |F_1(\xi)|^2 &\le \frac{1}{(1+\xi^2)^2\,(x^2+y^2)^2}, \\ |F_2(\xi)|^2 &\le \frac{64}{\left[x^2\,(1+\xi^2) + y^2\,(1+2\xi^2)^2\right]^2} \le \frac{64}{(1+\xi^2)^2\,(x^2+y^2)^2}. \end{aligned}$$

Finally,

$$|I_1| \le \frac{9\pi+2}{2(x^2+y^2)}. \tag{3.19}$$

We now turn to I_2. By integration by parts,

$$I_2 = -\int_0^\infty \frac{d}{d\xi}\left[H_2(\xi)\right] e^{K(\xi)} d\xi$$

where

$$H_2(\xi) = \frac{\xi^2 \left[ix - 4|y| \left(1+\xi^2\right)^{1/2}\right]}{\left[ix\left(1+\xi^2\right)^{1/2} - |y|\left(1+2\xi^2\right)\right]^3},$$

$$\begin{aligned} H_2'(\xi) = {} & \frac{2\xi\left[ix\left(1+\xi^2\right)^{1/2} - 2|y|\left(2+3\xi^2\right)\right]}{\left(1+\xi^2\right)^{1/2}\left[ix\left(1+\xi^2\right)^{1/2} - |y|\left(1+2\xi^2\right)\right]^3} \\ & - \frac{3\xi^3\left[ix - 4|y|\left(1+\xi^2\right)^{1/2}\right]^2}{\left(1+\xi^2\right)^{1/2}\left[ix\left(1+\xi^2\right)^{1/2} - |y|\left(1+2\xi^2\right)\right]^4} = G_1(\xi) + G_2(\xi). \end{aligned}$$

One easily checks that

$$\begin{aligned} |G_1(\xi)|^2 &\le \frac{64}{\left[x^2\left(1+\xi^2\right) + y^2\left(1+2\xi^2\right)^2\right]^2} \le \frac{64}{\left(1+\xi^2\right)^2\left(x^2+y^2\right)^2}, \\ |G_2(\xi)|^2 &\le \frac{9 \times 16^2}{\left(1+\xi^2\right)^2\left(x^2+y^2\right)^2}, \end{aligned}$$

and

$$|I_2| \le \frac{28\pi}{x^2+y^2}. \tag{3.20}$$

This leads to the estimate

$$|k_1(x,y)| \le \frac{C}{x^2+y^2}, \quad \text{if } y \neq 0. \tag{3.21}$$

In a similar way one obtains

$$|k_2(x,y)| \le \frac{C}{x^2+y^2} \quad \text{for } y \neq 0, \tag{3.22}$$

where

$$k_2(x,y) = -\int_{-\infty}^{0} \frac{\xi}{\left(1+\xi^2\right)^{1/2}} e^{|y|\xi\left(1+\xi^2\right)^{1/2} + ix\xi} d\xi.$$

It remains to consider the case where $y = 0$. But

$$k(x,0) = \int_{-\infty}^{\infty} \frac{|\xi|}{\left(1+\xi^2\right)^{1/2}} e^{ix\xi} d\xi,$$

and showing that $x^2 k(x,0)$ is bounded amounts to proving that $\mathcal{F}_x^{-1}[\frac{d^2}{d\xi^2}R(\xi)] \in L^\infty(\mathbb{R})$, where $R(\xi) = \frac{|\xi|}{(1+\xi^2)^{1/2}}$. A simple computation shows that

$$R''(\xi) = 2\delta\left(1+\xi^2\right)^{-1/2} + g(\xi) = 2\delta + g(\xi),$$

where $g \in L^1(\mathbb{R})$; this proves our claim. The proof of Lemma 3.6 is complete. □

We are now ready to prove (3.1). From (3.10) one has

(3.23)
$$\begin{aligned}\left|r^2(x,y)u(x,y)\right| \leq C &\left|\int_{\mathbb{R}^2} r^2(x-x',y-y')k(x-x',y-y')u^{p+1}(x',y')dx'dy'\right| \\ &+ C\left|\int_{\mathbb{R}^2} k(x-x',y-y')r^2(x',y')u^{p+1}(x',y')dx'dy'\right|.\end{aligned}$$

By Lemma 3.6 and the fact that $u \in L^q(\mathbb{R}^2)$, $q \geq 2$, the first term in the right-hand side of (3.23) is bounded independently of x and y. The second term is majorized by $\|k\|_{L^q}\left\|r^2u^{p+1}\right\|_{L^{q'}}$, $\frac{1}{q}+\frac{1}{q'}=1$. Recall that $k \in L^q(\mathbb{R}^2)$, $1<q\leq 2$ (see the proof of Lemma 3.5).

If $p>1$, we write

$$\left\|r^2u^{p+1}\right\|_{L^{q'}} \leq \|ru\|^2_{L^\infty}\left(\int_{\mathbb{R}^2} u^{q'(p-1)}\right)^{1/q'},$$

which is finite for some $q' \geq 2$.

The case $p=1$ needs a different argument. One has $|\nabla(ru)| \leq r|\nabla u| + |u|$. Let $s \in (1,2)$. Then, for $\varepsilon > 0$ small,

$$\int_{\mathbb{R}^2} r^s|\nabla u|^s = \int_{\mathbb{R}^2} r^s|\nabla u|^{\frac{2s}{4-\varepsilon}}|\nabla u|^{\frac{s(2-\varepsilon)}{4-\varepsilon}},$$

and by the Hölder inequality

$$\begin{aligned}\int_{\mathbb{R}^2} r^s|\nabla u|^s &\leq \left(\int_{\mathbb{R}^2} r^{4-\varepsilon}|\nabla u|^2\right)^{\frac{s}{4-\varepsilon}}\left(\int_{\mathbb{R}^2}|\nabla u|^{\frac{s(2-\varepsilon)}{4-\varepsilon-s}}\right)^{\frac{4-\varepsilon-s}{4-\varepsilon}} \\ &\leq C\left(\int_{\mathbb{R}^2}|\nabla u|^{\frac{s(2-\varepsilon)}{4-\varepsilon-s}}\right)^{\frac{4-\varepsilon-s}{4-\varepsilon}},\end{aligned}$$

by Lemma 3.5. For, say, $s=2-\varepsilon$, one has $\frac{s(2-\varepsilon)}{4-\varepsilon-s}<2$. On the other hand, $u, \nabla u \in L^q(\mathbb{R}^2)$, $1<q\leq 2$, as is easily checked from the identity $u=-k*\frac{u^2}{2}$ and the fact that $k \in L^q(\mathbb{R}^2)$, $1<q\leq 2$. Finally, $\nabla(ru) \in L^{2-\varepsilon}(\mathbb{R}^2)$, and by the Sobolev embedding theorem, $ru \in L^{\frac{2(2-\varepsilon)}{\varepsilon}}(\mathbb{R}^2)$, i.e., $r^2u^2 \in L^{q'}$, $q'=\frac{2-\varepsilon}{\varepsilon}>2$ for ε sufficiently small, and this proves that $r^2u \in L^\infty(\mathbb{R}^2)$. □

In the three-dimensional case we have the slightly less precise result, which follows.

Theorem 3.2. *Any nontrivial solitary wave of* (1.2) *satisfies*

$$r^\delta u \in L^2(\mathbb{R}^3)\ \forall\ \delta, 0 \leq \delta < 3/2,\quad r=(x^2+y^2+z^2)^{1/2}. \tag{3.24}$$

Proof. As in Lemma 3.1, one first proves that any solitary wave of (1.2) satisfies

$$\int_{\mathbb{R}^3}(x^2+y^2+z^2)(|\nabla u|^2+u^2_{xx})dxdydz < +\infty. \tag{3.25}$$

We still use (3.8), now with

$$\hat{h}(\xi_1,\xi_2,\xi_3)=\frac{\xi_1}{|\xi|^2+\xi_1^4},\quad |\xi|^2=\xi_1^2+\xi_2^2+\xi_3^2.$$

The following lemma is the counterpart of Lemma 3.2.

LEMMA 3.7.

$$rh \in L^q(\mathbb{R}^3), \quad 3 < q < 5, \quad r = (x^2+y^2+z^2)^{1/2}.$$

Proof. It suffices to prove that $\nabla_\xi \hat{h} \in L^{q'}(\mathbb{R}^3)$, $\frac{5}{4} < q' < \frac{3}{2}$. Since $\left|\nabla_\xi \hat{h}\right| \leq \frac{C}{|\xi|^2+\xi_1^4} = g(\xi)$, we are reduced to proving that $g \in L^{q'}(\mathbb{R}^3)$, $\frac{5}{4} < q' < \frac{3}{2}$. One finds readily that

$$\begin{aligned}\int_{\mathbb{R}^3} \frac{d\xi}{(|\xi|^2+\xi_1^4)^{q'}} &= \int_{\mathbb{R}} \frac{1}{|\xi_1|^{2q'}(1+\xi_1^2)^{q'}} \left[\int_{\mathbb{R}^2} \frac{d\xi_1 d\xi_2}{\left(1+\frac{\xi_2^2+\xi_3^2}{\xi_1^2\left(1+\xi_1^2\right)}\right)^{q'}}\right] d\xi_1 \\ &= \int_{\mathbb{R}} \frac{d\xi_1}{|\xi_1|^{2(q'-1)}(1+\xi_1^2)^{q'-1}} \int_{\mathbb{R}^2} \frac{d\xi_2 d\xi_3}{(1+\xi_2^2+\xi_3^2)^{q'}}.\end{aligned}$$

The first integral in the right-hand side is finite if and only if $\frac{5}{4} < q' < \frac{3}{2}$, and the second one, if and only if $q' > 1$. □

LEMMA 3.8.

$$ru \in L^\infty(\mathbb{R}^3).$$

Proof. From (3.8) we derive

$$|ru| \leq C\,|[rh] * u^p u_x| + C\,|h * [u^p r u_x]|\,.$$

By Young's inequality and Lemma 3.7, we have

$$\|(rh) * u^p u_x\|_{L^\infty} \leq \|rh\|_{L^4} \|u\|_{L^{4p}}^p \|u_x\|_{L^2} < +\infty.$$

On the other hand, it is easily seen that $\hat{h}$, hence h, belongs to $L^2(\mathbb{R}^3)$, and Young's inequality, together with (3.25), implies

$$\|h * u^p r u_x\|_{L^\infty} \leq \|h\|_{L^2} \|ru_x\|_{L^2} \|u\|_{L^\infty}^p < +\infty. \qquad \square$$

We now use equation (3.10), still valid in dimension 3, if we set

$$\hat{k}(\xi_1,\xi_2,\xi_3) = \frac{\xi_1^2}{|\xi|^2+\xi_1^4}.$$

LEMMA 3.9. *$\hat{k}$ belongs to the homogeneous Sobolev space $\dot{H}_2^s(\mathbb{R}^3)$, for any s with $\frac{7}{10} < s < \frac{3}{2}$.*

Proof. One checks easily that the second derivatives of $\hat{k}$ are bounded by $\frac{C}{|\xi|^2+\xi_1^4}$. Thus (see the proof of Lemma 3.7) $\partial^2_{\xi_i\xi_j}\hat{k} \in L^q(\mathbb{R}^3)$, $i,j = 1,2,3$, $\frac{5}{4} < q < \frac{3}{2}$; that is, $\hat{k}$ belongs to the homogeneous Sobolev space $\dot{H}_q^2(\mathbb{R}^3)$. Again from [2], we infer that $\dot{H}_q^2(\mathbb{R}^3) \subset \dot{H}_2^s(\mathbb{R}^3)$ for $s = \frac{7}{2} - \frac{3}{q}$, i.e., $\frac{11}{10} < s < \frac{3}{2}$.

On the other hand, a simple computation also shows that the first derivatives of $\hat{k}$ are bounded by $C|\hat{h}|$ and, by using the same argument as in the proof of Lemma 3.7, it is easily seen that $\hat{h} \in L^q(\mathbb{R}^3)$, $\frac{5}{3} < q < 2$; hence, reproducing the above reasoning yields $\hat{k} \in \dot{H}^1_q(\mathbb{R}^3) \subset \dot{H}^s_2(\mathbb{R}^3)$ for $s = \frac{5}{2} - \frac{3}{q}$ and $\frac{5}{3} < q < 2$, i.e., $\frac{7}{10} < s < 1$. The lemma follows by interpolation. □

We are now in position to conclude the proof of Theorem 3.2. We use the pointwise estimate (3.13). For any δ with $\frac{7}{10} < \delta < \frac{3}{2}$, one obtains, thanks to Lemma 3.9,

$$\left\|(r^\delta k) * u^{p+1}\right\|_{L^2} \leq C \left\|\hat{k}\right\|_{\dot{H}^\delta_2} \|u\|^{p+1}_{L^{p+1}} < +\infty. \tag{3.26}$$

On the other hand, denoting by $\mathcal{F}$ the Fourier transform,

$$\begin{aligned} \left\|k * (r^\delta u^{p+1})\right\|_{L^2} &= \left\|\mathcal{F}(k * r^\delta u^{p+1})\right\|_{L^2} = \left\|\hat{k}\mathcal{F}(r^\delta u^{p+1})\right\|_{L^2} \\ &\leq C \left\|r^\delta u^{p+1}\right\|_{L^2} \end{aligned} \tag{3.27}$$

since $\hat{k} \in L^\infty(\mathbb{R}^3)$. But

$$\left|\int_{\mathbb{R}^3} r^{2\delta} u^{2(p+1)}\right| \leq \|ru\|^{2\delta}_{L^\infty} \int_{\mathbb{R}^3} |u|^{p+2(1-\delta)}.$$

Now, $u \in L^q(\mathbb{R}^3)$ for any $q > 1$, as can be checked from (3.10) and the fact that (by Lizorkin's Theorem [11]), $\hat{k}$ is a Fourier multiplier in L^q, $1 < q < \infty$ (see [5]). Since $p \geq 1$, we conclude that $r^\delta u^{p+1} \in L^2(\mathbb{R}^3)$ for any $\delta < 1$. Together with (3.26), this implies that

$$r^\delta u \in L^2(\mathbb{R}^3), \quad \frac{7}{10} < \delta < 1. \tag{3.28}$$

Let us finally prove that (3.28) is also true for $\delta < \frac{3}{2}$. Let $\delta = \frac{3}{2} - \frac{\varepsilon}{2}$, $\varepsilon > 0$ small. Then

$$\left|\int_{\mathbb{R}^3} r^{2\delta} u^{2(p+1)}\right| \leq \left\|r^{1-\varepsilon} u\right\|_{L^2} \|ru\|^2_{L^\infty} \left\|u^{2p-1}\right\|_{L^2},$$

and the right-hand side is finite by (3.28) and Lemma 3.8. □

Remarks 3.2.

1. One proves similarly to (3.12) that $r^{1+\delta}\nabla u$, $r^{1+\delta}u_{xx} \in L^2(\mathbb{R}^3)$ for any δ, $0 \leq \delta < \frac{3}{2}$.

2. We do not know whether or not $r^3 u \in L^\infty(\mathbb{R}^3)$. Note that the assertion corresponding to (3.17) (that is, $r^3 k \in L^\infty(\mathbb{R}^3)$) is not true.

3. It is worth noting that all the results of sections 2 and 3, as those of [5], are valid mutatis mutandi for the BBM version of the generalized KP equations considered here, namely, when the u_{xxx} term in (1.1) or (1.2) is replaced by $-u_{xxt}$.

4. An extension. A natural question is whether or not the results in Chapter 3 are modified by adding to (1.1) or (1.2) a higher order dispersive term in x. Essentially, they are not. To keep this paper short we will restrict ourselves to the two-dimensional and three-dimensional versions of a fifth-order KdV equation introduced by Abramyan and Stepanyants [1] and Karpman and Belashov [9], [10].

$$\begin{cases} u_t + u^p u_x + u_{xxx} + \delta u_{xxxxx} - v_y = 0, \\ v_x = u_y \end{cases} \tag{4.1}$$

in the 2-dimensional case and

$$\begin{cases} u_t + u^p u_x + u_{xxx} + \delta u_{xxxxx} - v_y - w_z = 0, \\ v_x = u_y, \quad w_x = u_z \end{cases} \tag{4.2}$$

in the 3-dimensional case.

A solitary wave of (4.1) (resp., (4.2)) is a solution of the form $u(x - ct, y)$ (resp., $u(x - ct, y, z)$) where $c > 0$ and $u \in Z = \{u \in Y,\ \partial_x^2 u \in L^2(\mathbb{R}^d)\}$, $d = 2, 3$.

In [5], we proved that when $\delta = -1$ (which we will assume from now on), (4.1) (resp., (4.2)) has a nontrivial solitary wave for arbitrary p's (resp., $1 \le p < \frac{8}{3}$), which belongs to $H^\infty(\mathbb{R}^d)$, $d = 2, 3$, when p is an integer.

Concerning the properties of such solitary waves, we have the following theorem.

THEOREM 4.1.

(i) *The solitary waves of* (4.1) *satisfy*

$$r^2 u \in L^\infty(\mathbb{R}^2), \quad r^2 = x^2 + y^2.$$

(ii) *The solitary waves of* (4.2) *satisfy*

$$r^\delta u \in L^2(\mathbb{R}^3), \quad r^2 = x^2 + y^2 + z^2, \ \textit{for any } \delta,\ 0 \le \delta < \frac{3}{2}.$$

The proof of Theorem 4.1 follows, with some technical differences, the corresponding ones in sections 2 and 3 and will be omitted. Note that the decay estimates are sharp (Remark 3.1 is still valid in this context); they are essentially imposed by the singularity of $\hat{k}$ at 0, namely, $\hat{k}(\xi) \sim \frac{\xi_1^2}{|\xi|^2}$.

Appendix. We state here the unique continuation result we have used in section 2. We consider first the two-dimensional situation.

THEOREM A.1. *Let $a, b, c \in L^\infty(\mathbb{R}^2)$ and u satisfy*

$$u, u_y, u_{xy}, u_{xx}, u_{xxx} \in L^2(\mathbb{R}^2), \tag{A.1}$$

$$u_{yy} - u_{xxxx} = a(x,y)u + b(x,y)u_x + c(x,y)u_{xx} \ \textit{in } \mathbb{R}^2. \tag{A.2}$$

Then, if u vanishes on a half-plane Π in $\mathbb{R}^2$, it vanishes everywhere in $\mathbb{R}^2$.

Proof. If the line delimiting Π is not characteristic (that is, not parallel to the x-axis), the result is an easy consequence of Isakov's unique continuation theorem [8]. If $\partial\Pi$ is parallel to the x-axis (which is the case we need in section 2), we have to use a different (global) argument which we only sketch here. It suffices obviously to prove that if u satisfying (A.1) and (A.2) is such that $u \equiv 0$ on $\{(x,y), y \le 0\}$ then it vanishes on $\Pi_T = \{(x,y), 0 \le y \le T\}$ for any $T > 0$.

We write (A.2) as

$$\frac{\partial v}{\partial y} - Av = -v + au + bu_x + cu_{xx}, \tag{A.3}$$

where

$$v = u_y - u_{xx},$$

$$A = I - \frac{\partial^2}{\partial x^2}.$$

A is a self-adjoint operator from $V = H^1(\mathbb{R})$ to $H^{-1}(\mathbb{R})$, which satisfies the hypothesis of Proposition II.1 in [15]. In order to apply the backward uniqueness result (Theorem II.1) of [15] (note that the "time" variable y is reversed in our problem), proving that v vanishes on Π_T, it suffices that the right-hand side of (A.3) defines an operator $B(y) \in L^2_y(0, T, \mathcal{L}(H^1_x(\mathbb{R}), L^2_x(\mathbb{R})))$. This is easily checked by using (A.1) and the fact that u solves the heat equation in Π_T

$$\begin{cases} u_y - u_{xx} = v, \\ u(x, 0) = 0. \end{cases} \tag{A.4}$$

Then, by the uniqueness of the Cauchy problem (A.4), u vanishes on Π_T.

For the three-dimensional situation we state the following theorem.

THEOREM A.2. *Let $a, b, c \in L^\infty(\mathbb{R}^3)$ and u satisfy*

$$u, u_y, u_z, u_{xy}, u_{xz}, u_{yz}, u_{yy}, u_{zz}, u_{xxx}, u_{xxy}, u_{xxz} \in L^2(\mathbb{R}^3), \tag{A.5}$$

$$u_{yy} + u_{zz} - u_{xxxx} = a(x, y, z)u + b(x, y, z)u_x + c(x, y, z)u_{xx}. \tag{A.6}$$

Then, if u vanishes on one side of a hyperplane H, it vanishes everywhere in $\mathbb{R}^3$.

Proof. Again, if H is not characteristic (that is, not parallel to the x-axis) the result follows from Isakov's theorem [8]. If H is parallel to the x-axis (which is the case of interest in section 2), we use a global argument. First, by the invariance of $\partial_y^2 + \partial_z^2$ by rotations in the (y, z) plane, it suffices to consider the case where H is parallel to a plane of coordinates containing the x-axis; for instance, $H = \{(x, y, z), y = 0\}$. Let us assume that u vanishes on $\{(x, y, z), y \leq 0\}$ and let us prove that u vanishes on

$$\Pi_T = \{(x, y, z), 0 \leq y \leq T\}$$

for any $T > 0$. We factorize (A.6) as

$$\frac{\partial v}{\partial y} - \mathcal{A}v = -v + au + bu_x + cu_{xx}, \tag{A.7}$$

where

$$v = u_y + Au, \mathcal{A} = I + A,$$

A being the operator (in $\mathbb{R}^2$) defined in Fourier variables by

$$\widehat{Au}(\xi_1, \xi_3) = (\xi_3^2 + \xi_1^4)^{1/2} \widehat{u}(\xi_1, \xi_3).$$

Obviously, A is a self-adjoint operator, continuous from V into V', where

$$V = \{v \in L^2(\mathbb{R}^2), v_x \in L^2(\mathbb{R}^2), |\xi_3|^{1/2}\widehat{v} \in L^2(\mathbb{R}^2)\}.$$

Again, it is easily checked that the hypothesis of Theorem II.1 in [15] is satisfied, yielding $v \equiv 0$ on Π_T and, therefore, $u \equiv 0$ on Π_T.

Acknowledgment. We thank Thierry Colin, who suggested that we use [12] to obtain the symmetry of the solitary waves.

REFERENCES

[1] L. A. Abramyan and Y. A. Stepanyants, *The structure of two-dimensional solitons in media with anomalously small dispersion*, Sov. Phys. JETP, 61 (1985), pp. 963–966.

[2] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, Berlin, 1976.

[3] O. V. Besov, V. P. Il'in, and S. M. Nikolskii, *Integral Representations of Functions and Imbeddings Theorems*, Vol. I, Wiley, New York, 1978.

[4] J. L. Bona and Yi A. Li, *Decay and analyticity of solitary waves*, preprint, 1995.

[5] A. de Bouard and J.C. Saut, *Solitary waves of the generalized Kadomtsev–Petviashvili equations*, Annales Institut Henri Poincaré, Analyse Non Linéaire, 4 (1997), pp. 211–236.

[6] T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, Textos de Métodos Matematicos 26, Instituto de Matematica-UFRJ Rio de Janeiro, Brazil, 1993.

[7] T. Colin and M. Weinstein, *On the ground states of vector nonlinear Schrödinger equations*, preprint, 1995.

[8] V. Isakov, *Carleman type estimates in an anisotropic case and applications*, J. Differential Equations, 105, (1993), pp. 217–238.

[9] V. I. Karpman and V. Yu. Belashov, *Dynamics of two-dimensional solitons in weakly dispersive media*, Phys. Lett. A, 154 (1991), pp. 131–139.

[10] V. I. Karpman and V. Yu. Belashov, *Evolution of three-dimensional nonlinear pulses in weakly dispersive media*, Phys. Lett. A, 154 (1991), pp. 140–144.

[11] P. I. Lizorkin, *Multipliers of Fourier integrals*, Proc. Steklov Inst. Math., 89 (1967), pp. 269–290.

[12] O. Lopes, *A constrained minimization problem with integrals on the entire space*, Bol. Soc. Bras. Mat., 25 (1994), pp. 77–92.

[13] J. C. Saut, *Remarks on the generalized Kadomtsev–Petviashvili equations*, Indiana Math. J., 42 (1993), pp. 1011–1026.

[14] X. P. Wang, M. J. Ablowitz, and H. Segur, *Wave collapse and instability of solitary waves for a generalized Kadomtsev–Petviashvili equation*, Phys. D, 78 (1994), pp. 97–113.

[15] C. Bardos and L. Tartar, *Sur l'unicité rétrograde des équations paraboliques et quelques questions voisines*, Arch. Rational Mech. Anal., 50 (1973), pp. 10–25.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1086–1093, September 1997

A NOTE ON A TWO-POINT BOUNDARY VALUE PROBLEM ARISING FROM A LIQUID METAL FLOW*

YONGDONG SHI†, QINDE ZHOU†, AND YONG LI†

Abstract. A two-point boundary value problem with a positive parameter Q arising in the study of surface-tension–induced flows of a liquid metal or semiconductor is studied. On the basis of the upper–lower solution method and Schauder's fixed-point theorem, it is proved that when $0 \leq Q \leq 13.213$, the problem admits a solution. This improves a recent result where $0 \leq Q < 1$.

Key words. two-point boundary value problem, the upper–lower solution method, Schauder's fixed-point theorem

AMS subject classification. 34B

PII. S0036141095290252

1. Introduction.

Consider the following nonautonomous two-point boundary value problem (BVP) on [0,1]:

$$\begin{cases} \left[x\left(\frac{f'}{x}\right)'\right]' + Q\left[f\left(\frac{f'}{x}\right)' - x\left(\frac{f'}{x}\right)^2\right] = \beta x, & \text{(1.1a)} \\ f(0) = f(1) = \left(\frac{f'}{x}\right)'\Big|_{x=0} = \left(\frac{f'}{x}\right)'\Big|_{x=1} - 1 = 0, & \text{(1.1b)} \end{cases}$$

where $' = d/dx$. This problem arises in the study of surface-tension–induced flows of a liquid metal or semiconductor in a cylindrical floating zone of length $2L$ and radius R. Here the parameter $Q = 2L^3R^{-3}(\text{Re})$, Re is the Reynolds number, and β is a constant to determine.

Numerical solutions of (1.1) have been found [1] for $0 \leq Q \leq 32.7$ and $Q \geq 1749$. However, a theoretical proof of the existence of solutions of (1.1) has been done only for $0 \leq Q < 1$ in [2]. Hence, there is still a large gap between numerical experiments and theoretical results. In the present paper, on the basis of the upper–lower solution method and Schauder's fixed-point theorem, we prove the existence of solutions for (1.1) with $0 \leq Q \leq 13.213$. Thereby we greatly improve the existing results [2].

Our main result is the following.

THEOREM 1. *For* $0 \leq Q \leq 13.213$, *there exists a constant* β *such that* (1.1) *admits a solution* $f = f(x)$ *satisfying, on* $(0, 1)$,

$$-0.039064 \leq f(x) \leq 0, \quad -0.089647 \leq f'(x) \leq \frac{1}{3}, \quad -0.388320 \leq f''(x) \leq \frac{4}{3}.$$

2. A technical treatment of (1.1).

We observe that in (1.1), (1.1a) is a third-order equation with an unknown constant β, while the boundary value condition (1.1b) contains four equalities. Hence, following [2], we make the following technical treatment of (1.1).

[*] Received by the editors August 11, 1995; accepted for publication (in revised form) April 11, 1996.

http://www.siam.org/journals/sima/28-5/29025.html

[†] Department of Mathematics, Jilin University, Changchun 130023, People's Republic of China (math@mail.jlu.edu.cn).

Differentiating (1.1a) with respect to x, we obtain

$$\begin{cases} \left[\left(\frac{f'}{x}\right)'\right]'' + \left[\frac{1+Qf}{x}\right]\left(\frac{f'}{x}\right)'' - \left[\frac{1+Q(xf)'}{x^2}\right]\left(\frac{f'}{x}\right)' = 0, & \text{(2.1a)} \\ f(0) = f(1) = \left(\frac{f'}{x}\right)'\Big|_{x=0} = \left(\frac{f'}{x}\right)'\Big|_{x=1} - 1 = 0. & \text{(2.1b)} \end{cases}$$

Let $(f'/x)' = g$. Then (2.1) has the form

$$\begin{cases} g'' + \left[\frac{1+Qf}{x}\right] g' - \left[\frac{1+Q(xf)'}{x^2}\right] g = 0, & \text{(2.2a)} \\ g(0) = g(1) - 1 = 0. & \text{(2.2b)} \end{cases}$$

To prove the existence of solutions for (1.1), we reduce it to a problem of finding a fixed point. On the basis of the differential inequality technique, to construct upper and lower solutions of (2.2), we consider the following set:

$$D = \{f | f \in C^1[0,1], f(0) = f(1) = 0, h(x) \le f(x) \le 0, \\ m(x) \le f'(x) \le n(x)\},$$

where

$$h(x) = \frac{(25)^2}{3219} x^2 (x^{\frac{37}{25}} - 1), \qquad n(x) = \frac{1}{3} x^4,$$

$$m(x) = 25x \left(\frac{1}{126} x^{\frac{126}{25}} - \frac{1}{87} x^{\frac{87}{25}} + \frac{1}{37} x^{\frac{37}{25}} - \frac{50}{3219} \right).$$

For any $f \in D$, if (1) eq. (2.2) has a unique solution $g(x)$ and (2) the problem

$$\begin{cases} \left(\frac{f^{*\prime}}{x}\right)' = g, & \text{(2.3a)} \\ f^*(0) = f^*(1) = 0 & \text{(2.3b)} \end{cases}$$

also has a unique solution $f^*(x)$, then we may define an operator

$$T : f \longmapsto f^*, f \in D,$$

where f^* is the solution of (2.3). Thus, given $Q \in [0, 13.213]$, if we can prove that (3) T has a fixed point, namely, there exists $f \in D$ such that $Tf = f$, then f is a solution of (2.1). Integrating (2.1a) from 1 to x and using (2.1b), we obtain (1.1a) at once; here $\beta = [(\frac{f'}{x})'' + \frac{1+Qf}{x}(\frac{f'}{x})' - Q(\frac{f'}{x})^2]\big|_{x=1}$. Therefore, f must be the solution of (1.1).

In the following, we shall carry out the above three processes, respectively.

3. The solution of the problem (2.2). We consider the boundary value problem on $[x_1, x_2]$

$$y'' = a(x)y' + b(x)y, \tag{3.1a}$$

$$y(x_1) = A_1, y(x_2) = A_2, \tag{3.1b}$$

where $a(x), b(x) \in C^1[x_1, x_2]$ and $b(x) > 0$.

LEMMA 1. *Suppose that there exist functions* $\overline{\omega}(x), \underline{\omega}(x) \in C^2[x_1, x_2]$ *such that for* $x_1 \leq x \leq x_2$,

$$\begin{aligned}&\underline{\omega}(x) \leq \overline{\omega}(x),\\&\overline{\omega}''(x) \leq a(x)\overline{\omega}'(x) + b(x)\overline{\omega}(x),\\&\underline{\omega}''(x) \geq a(x)\underline{\omega}'(x) + b(x)\underline{\omega}(x),\end{aligned}$$

and

$$\underline{\omega}(x_i) \leq A_i \leq \overline{\omega}(x_i), \qquad i = 1, 2.$$

Then the problem (3.1) *has a unique solution* $y = y(x)$, *and*

$$\underline{\omega}(x) \leq y(x) \leq \overline{\omega}(x), \qquad x_1 \leq x \leq x_2.$$

Moreover, there exists a positive number N *which depends only on interval* $[x_1, x_2]$ *and the function pairs* $\overline{\omega}(x), \underline{\omega}(x)$ *such that*

$$|y'(x)| \leq N, \qquad x_1 \leq x \leq x_2.$$

Since $b(x) > 0$, we use the maximal value principle, it is easy to prove the uniqueness of solutions of (3.1), and other aspects of Lemma 1 are generalizations of Nagumo's theorem (see [3, Thm. 1.5.1]).

THEOREM 2. *Assume* $f \in D$ *and* $0 \leq Q \leq 13.213$. *Then the boundary value problem* (2.2) *has a unique solution* $g = g(x)$.

Proof. Notice for $f \in D, x \in (0, 1]$, we have

$$\begin{aligned}\frac{1 + Q(xf)'}{x^2} = \frac{1}{x^2}[1 + Q(xf' + f)] &\geq \frac{1}{x^2}[1 + Q(xm(x) + h(x))]\\&= \frac{1}{x^2}[1 + QF(x)],\end{aligned}$$

where

$$F(x) = xm(x) + h(x) = 25x^2\left(\frac{1}{126}x^{\frac{126}{25}} - \frac{1}{87}x^{\frac{87}{25}} + \frac{112}{3219}x^{\frac{37}{25}} - \frac{75}{3219}\right).$$

By a simple argument, we show that $F(x)$ is decreasing on $[0, c]$ and increasing on $[c, 1]$, where $c = 0.565711027\ldots$. Hence $F(x)$ takes a minimum at $x = c$, i.e.,

$$\min_{x\in[0,1]} F(x) = F(c) = -0.0756773788\cdots \geq -0.075678.$$

Therefore, for $0 \leq Q \leq 13.213$, we have

$$\frac{1 + Q(xf)'}{x^2} > 0, \qquad x \in (0, 1]. \tag{3.2}$$

For any positive integer $n \geq 2$, consider the boundary value problem

$$\begin{cases} g'' = -\left[\dfrac{1 + Qf}{x}\right]g' + \left[\dfrac{1 + Q(xf)'}{x^2}\right]g, & \text{(3.3a)}\\ g(\frac{1}{n}) = 0, \qquad g(1) = 1. & \text{(3.3b)}\end{cases}$$

We set $\underline{\omega}(x) \equiv 0$, $\overline{\omega}(x) = x^\alpha$, where $\alpha > 0$ is sufficiently small, so that

$$\alpha^2 + \alpha Q f \leq 1 + Q(xf)'.$$

Then

$$\begin{aligned}\overline{\omega}''(x) &\leq -\left[\frac{1+Qf}{x}\right]\overline{\omega}'(x) + \left[\frac{1+Q(xf)'}{x^2}\right]\overline{\omega}(x),\\ \underline{\omega}''(x) &\geq -\left[\frac{1+Qf}{x}\right]\underline{\omega}'(x) + \left[\frac{1+Q(xf)'}{x^2}\right]\underline{\omega}(x)\end{aligned}$$

for all $x \in [\frac{1}{n}, 1]$, $n \geq 2$, $\underline{\omega}(1) = 0 < 1 = \overline{\omega}(1)$, and $\underline{\omega}(\frac{1}{n}) = 0 < \frac{1}{n^\alpha} = \overline{\omega}(\frac{1}{n})$. By Lemma 1, we obtain that (3.3) has only one solution, $g_n = g_n(x)$, which satisfies

$$0 \leq g_n(x) \leq x^\alpha, \qquad x \in \left[\frac{1}{n}, 1\right].$$

Since $g_n(x), n = 2, 3, \ldots$ are all solutions of (3.3a) and $o \leq g_n(x) \leq x^\alpha, x \in [\frac{1}{2}, 1]$, applying the part related to the estimate of the derivative in Lemma 1, we have that $\{g_n'(x)\}$ is uniformly bounded on $[\frac{1}{2}, 1]$ and hence that $\{g_n'(1)\}$ is bounded. Without loss of generality, we let $\{g_n'(1)\} \to \alpha_0$ as $n \to \infty$.

We consider the solution of (3.3a) satisfying the initial conditions $g(1) = 1$, $g'(1) = \alpha_0$. Obviously, it exists on $[0, 1]$ and satisfies

$$0 \leq g(x) \leq x^\alpha;$$

namely, $g(x)$ is the solution of (2.2). The uniqueness of the solution is easy to obtain by (3.2). The proof of the theorem is completed.

To prove Theorem 1, we give the bound of $g(x)$ and $g'(x)$ on $[0, 1]$.

THEOREM 3. *For $f \in D$, $0 \leq Q \leq 13.213$, the solution $g(x)$ of* (2.2) *satisfies*

(3.4) (i) $g(x) > 0,\ \ g'(x) > 0, 0 < x < 1,$

(3.5) (ii) $\lim_{x \to 0+} x g'(x) = 0,$

(3.6) (iii) $x^{\frac{51}{25}} \leq g(x) \leq x^{\frac{12}{25}},$

(3.7) (iv) $\dfrac{12}{25} x^{\frac{51}{25}} \leq x g'(x) \leq \dfrac{51}{25} x^{\frac{12}{25}}.$

Proof. (i) Since $1+Q(xf)' > 0$, $x \in [0, 1]$, we see easily that $g(x) > 0$, $g'(x) \geq 0$ on $(0, 1)$ by the maximal value principle. Next we assert that $g'(x) \neq 0$ for any $x \in (0, 1)$. If not, then there exists $x_0 \in (0, 1)$ such that $g'(x_0) = 0$. We have $g''(x_0) > 0$ from (2.2a), namely, $g(x)$ takes a minimum at $x = x_0$. Hence $g(x)$ must have a maximum at some $x_1 \in (0, x_0)$. This is impossible because we have $g''(x_1) > 0$ from (2.2a), a contradiction.

(ii) Rewrite (2.2a) as follows:

$$x(xg')' = -Qfg' + [1 + Q(xf)']g.$$

Owing to $0 \leq x \leq 1$, $f \leq 0$, $1 + Q(xf)' > 0$, and $g > 0$, we have $x(xg')' > 0$; i.e., $(xg')' > 0$, and hence xg' is increasing on $(0, 1)$. Using (3.4), we obtain $xg' > 0$ for $x \in (0, 1)$, and therefore $\lim_{x \to 0+} xg'(x)$ exists and $\lim_{x \to 0+} xg'(x) \geq 0$. If there exists $\alpha > 0$ such that $\lim_{x \to 0+} xg'(x) = \alpha$, then for $\frac{\alpha}{2}$, there is a $\delta > 0$ so that

$\frac{\alpha}{2} < xg'(x) < g'(1)$ for $x \in (0, \delta)$, or $\frac{\alpha}{2x} < g'(x) < \frac{1}{x}g'(1)$. Integrating it from x to δ, we have

$$\frac{\alpha}{2} \ln \frac{\delta}{x} < g(\delta) - g(x) < g'(1) \ln \frac{\delta}{x}.$$

This means $g(x) \to -\infty$ as $x \to 0+$, contradicting $g(0) = 0$. Thus $\lim_{x\to 0+} xg'(x) = 0$.

(iii) Equation (2.2a) can be converted to the following form:

$$(x^2g')' - (xg)' - Q(xfg)' = -2Qxfg'.$$

Integrating the above equation from 0 to x, using (3.5) and $g(0) = 0$, we obtain

$$g' = \left[\frac{1+Qf}{x}\right] g - \frac{2Q}{x^2} \int_0^x tfg'\, dt, \qquad 0 < x < 1.$$

Hence, as $-0.039064 \le f(x) \le 0$, $g'(x) \ge 0$, and $0 \le Q \le 13.213$, we have $Qf > -\frac{13}{25}$. This implies

$$\begin{aligned} \frac{12}{25x}\, g(x) \le g'(x) \quad &\le \frac{1}{x}\, g(x) + \frac{26}{25x} \int_0^x g'\, dt \\ &= \frac{51}{25x}\, g(x). \end{aligned} \tag{3.8}$$

We integrate (3.8) from 1 to x and obtain (3.6).

(iv) We combine (3.6) with (3.8) and yield (3.7).

The theorem is proved.

4. The solution of boundary value problem (2.3). Integrating (2.3a) from 0 to x and using (2.3b), we see

$$f^{*\prime}(x) = kx + x \int_0^x g(t)\, dt \tag{4.1}$$

and

$$f^*(x) = \frac{1}{2}kx^2 + \int_0^x \left(s \int_0^s g(t)\, dt \right) ds, \tag{4.2}$$

where $k = -2\int_0^1 (s \int_0^s g(t)\, dt)\, ds$. By Theorem 2, we know that $g(x)$ exists and is unique, so (2.3) has a unique solution $f^* = f^*(x)$ on $[0, 1]$.

In the following, we estimate the bound of f^* and $f^{*\prime}$.

Since

$$k = -2 \int_0^1 \left(s \int_0^s g(t)\, dt \right) ds = -2 \int_0^1 \left(g(t) \int_t^1 s\, ds \right) dt = -\int_0^1 (1 - t^2) g(t)\, dt,$$

(4.1) and (4.2) become

$$f^{*\prime}(x) = -x \int_0^1 (1 - t^2) g(t)\, dt + x \int_0^x g(t)\, dt, \tag{4.3}$$

$$f^*(x) = -\frac{1}{2}x^2 \int_0^1 (1 - t^2) g(t)\, dt + \frac{1}{2} \int_0^x (x^2 - t^2) g(t)\, dt. \tag{4.4}$$

From (4.3), we have

$$\begin{aligned} f^{*\prime}(x) &= -x\int_0^x (1-t^2)g(t)\,dt - x\int_x^1 (1-t^2)g(t)\,dt + x\int_0^x g(t)\,dt \\ &= -x\int_x^1 (1-t^2)g(t)\,dt + x\int_0^x t^2 g(t)\,dt. \end{aligned}$$

By (3.4) and (3.6), we have the following inequalities:

$$\begin{aligned} f^{*\prime}(x) &\le x\int_0^x t^2\,dt = \frac{1}{3}x^4 = n(x) \le \frac{1}{3}, \\ f^{*\prime}(x) &\ge -x\int_x^1 (1-t^2)t^{\frac{12}{25}}\,dt + x\int_0^x t^2 t^{\frac{51}{25}}\,dt \\ &= 25x\left(\frac{1}{126}x^{\frac{126}{25}} - \frac{1}{87}x^{\frac{87}{25}} + \frac{1}{37}x^{\frac{37}{25}} - \frac{50}{3219}\right) \\ &= m(x) \ge -0.089647. \end{aligned}$$

From (4.4), it follows that

$$f^*(x) = \frac{1}{2}(x^2-1)\int_0^x t^2 g(t)\,dt - \frac{1}{2}x^2\int_x^1 (1-t^2)g(t)\,dt.$$

By (3.4) and (3.6), we obtain

$$\begin{aligned} 0 \ge f^*(x) &\ge \frac{1}{2}(x^2-1)\int_0^x t^2 t^{\frac{12}{25}}\,dt - \frac{1}{2}x^2\int_x^1 (1-t^2)t^{\frac{12}{25}}\,dt \\ &= \frac{(25)^2}{3219}x^2(x^{\frac{37}{25}}-1) \\ &= h(x) \ge -\left(\frac{25}{87}\right)^2\left(\frac{50}{87}\right)^{\frac{50}{37}} \ge -0.039064. \end{aligned}$$

Using similar arguments, we have

$$-0.388320 \le f^{*\prime\prime}(x) \le \frac{4}{3}.$$

In summary, we get the following inequalities:

$$-0.039064 \le f^*(x) \le 0,$$

$$-0.089647 \le f^{*\prime}(x) \le \frac{1}{3},$$

$$-0.388320 \le f^{*\prime\prime}(x) \le \frac{4}{3}.$$

5. *T* has a fixed point. We define the norm on $C^1[0,1]$ by

$$||f|| := \max|f| + \max|f'|, \qquad x \in [0,1].$$

Then $C^1[0,1]$ is a Banach space. It is easy to check that D in section 2 is a nonempty, closed, bounded, convex subset of $C^1[0,1]$. By sections 3 and 4, we see that T is well defined and $TD \subseteq D$. In addition, T maps a bounded subset of D into a compact subset of D. (For details, see [2].)

Now we prove that the operator T is continuous.

By the definition of T and (4.2), we only need prove that for any given $f_0 \in D$ and any $\varepsilon > 0$, there exists $\delta > 0$ such that as $||f - f_0|| < \delta$ and $f \in D$,

$$\max_{x\in[0,1]} |g(x) - g_0(x)| < \varepsilon, \tag{5.1}$$

where g and g_0 are solutions of (5.2) and (5.3), respectively:

$$\begin{cases} g'' + \left[\dfrac{1+Qf}{x}\right] g' - \left[\dfrac{1+Q(xf)'}{x^2}\right] g = 0, \\ g(0) = g(1) = 0, \qquad x \in [0,1]. \end{cases} \tag{5.2}$$

$$\begin{cases} g'' + \left[\dfrac{1+Qf_0}{x}\right] g' - \left[\dfrac{1+Q(xf_0)'}{x^2}\right] g = 0, \\ g(0) = g(1) = 0, \qquad x \in [0,1]. \end{cases} \tag{5.3}$$

Let $p(x) = g(x) - g_0(x)$. Then by (5.2) and (5.3), we have

$$\begin{aligned} L[p] &= p'' + [1+Qf_0]\, p' - \left[\frac{1+Q(xf_0)'}{x^2}\right] p \\ &= -\left(\tfrac{Q}{x^2}\right)[(f-f_0)(xg'-g) - (f'-f_0')xg] \\ &= -G(x), \end{aligned} \tag{5.4}$$

with $p(0) = p(1) = 0$. For any $0 \le x \le \varepsilon$, $f \in D$, and $0 \le Q \le 13.213$, by using (3.6) we obtain

$$|p(x)| = |g(x) - g_0(x)| \le 2x^{\frac{12}{25}} \le 2\varepsilon^{\frac{12}{25}} = \varepsilon_1. \tag{5.5}$$

For $x \in [\varepsilon, 1]$, $||f - f_0|| \to 0$, and $0 \le Q \le 13.213$, we claim that $|p(x)| < \varepsilon_1$.

Set

$$\varepsilon^* = (1 - 0.075678Q)\varepsilon_1. \tag{5.6}$$

If $||f - f_0||$ is sufficiently small, $x \in [\varepsilon, 1]$, and $0 \le Q \le 13.213$, then, by (3.6) and (3.7), we have

$$\begin{aligned} |G(x)| &\le \tfrac{Q}{\varepsilon^2}[\max\{|xg'| + |xg| + |g|\}]||f - f_0|| \\ &\le \tfrac{Q}{\varepsilon^2}(\tfrac{51}{25} + 1 + 1)||f - f_0|| \\ &= \tfrac{101Q}{25\varepsilon^2}||f - f_0|| < \varepsilon^*. \end{aligned} \tag{5.7}$$

For fixed f_0, let p be a solution of (5.4) with the boundary conditions $|p(\varepsilon)| \le \varepsilon_1$ and $p(1) = 0$. By (5.7), we have

$$L[p] - \varepsilon^* \le L[p] + F(x) \le L[p] + \varepsilon^*, \qquad \varepsilon \le x \le 1. \tag{5.8}$$

Let $p^{\pm}$ be solutions of the following problems:

$$\begin{cases} L[p] \pm \varepsilon^* = 0, \\ p^{\pm}(1) = 0, \qquad p^{\pm}(\varepsilon) = p(\varepsilon). \end{cases}$$

Then, using the comparison theorem, we show that

$$p^-(x) \le p(x) \le p^+(x), \qquad x \in [\varepsilon, 1].$$

Now we prove that for any $x \in [\varepsilon, 1]$, there is $p^+(\varepsilon) \le \varepsilon_1$. In fact, we see that $p^+(x) \le \varepsilon_1$ as $|p(\varepsilon)| \le \varepsilon_1$. If not, then there exists a point $x_+ \in (\varepsilon, 1)$ such that $p^+(x_+) > \varepsilon_1$. By $p^+(\varepsilon) = p(\varepsilon) \le \varepsilon_1$, there must be a point $y_+ \in (\varepsilon, 1)$ such that $p^+(x)$ takes maximum at y_+, namely,

$$p^+(y_+) > \varepsilon_1, \quad {p^+}'(y_+) = 0, \quad {p^+}''(y_+) < 0.$$

But by (5.6) and (5.9), we see that the following holds:

$$\begin{aligned} {p^+}''(y_+) + \varepsilon^* &= \left[\frac{1 + Qf_0(y_+) + Qf_0'(y_+)y_+}{y_+^2}\right] p^+(y_+) \\ &> (1 - 0.075678Q)\varepsilon_1. \end{aligned}$$

Hence, using (5.6), we get

$${p^+}''(y_+) > (1 - 0.075678Q)\varepsilon_1 - \varepsilon^* = 0,$$

a contradiction. For $x \in [\varepsilon, 1]$ and $|p(\varepsilon)| \le \varepsilon_1$, we argue similarly and obtain $p^-(x) \ge -\varepsilon_1$. Thus, when $x \in [\varepsilon, 1]$ and $|p(\varepsilon)| \le \varepsilon_1$, there exists $|p(x)| \le \varepsilon_1$.

Therefore, for any given $\varepsilon > 0$, if we choose

$$\delta = \frac{\varepsilon^*}{\frac{101Q}{25\varepsilon^2}},$$

where ε^* satisfies (5.6), then for $f \in D$, $\max_{x\in[0,1]} |g(x) - g_0(x)| = \varepsilon_1 = 2\varepsilon^{\frac{12}{25}}$ as $||f - f_0|| < \delta$, and the continuity of T is proved.

To sum up, we see that the operator T satisfies the conditions of Schauder's fixed-point theorem [4], and thus T has at least one fixed point in D.

By sections 2, 3, 4, and 5, Theorem 1 is proved.

Acknowledgment. The authors thank the referees for their valuable suggestions.

REFERENCES

[1] W. N. Gill, N. D. Kazarinoff, C. Hus, M. Noack, and J. D. Verhoeven, *Thermo-capillary driven convection in supported and floating zone crystallization*, Adv. Space Res., 4 (1984), pp. 15–22.

[2] C. Lu and N. D. Kazarinoff, *On the existence of solution of a two-point boundary value problem arising from flows in a cylindrical floating zone*, SIAM. J. Math. Anal., 20 (1989), pp. 494–503.

[3] S. R. Bernfeld and V. Lakshmikantham, *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1974.

[4] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1094–1112, September 1997

006

STABILITY OF TRAVELLING MULTIPLE-FRONT (MULTIPLE-BACK) WAVE SOLUTIONS OF THE FITZHUGH–NAGUMO EQUATIONS*

SHUNSAKU NII†

Abstract. Consideration is devoted to travelling multiple-front (back) wave solutions of the FitzHugh–Nagumo equations of bistable type. In particular, stability of the 1-front (back) wave is proven. In the proof, the eigenvalue problem for the 1-front wave bifurcating from coexisting simple front and back waves is regarded as a bifurcation problem for projectivized eigenvalue equations, rather than treated as a linear eigenvalue problem for each fixed wave.

Key words. travelling wave, eigenvalue problem, bifurcation

AMS subject classifications. 34, 35

PII. S003614109528829X

1. Introduction.

The following system is called the FitzHugh–Nagumo equations:

$$\begin{cases} u_t &= u_{xx} + f(u) - w, \\ w_t &= \varepsilon(u - \gamma w), \end{cases} \tag{1.1}$$

where $x, t \in \mathbb{R}$ and $u(x,t), w(x,t) \in \mathbb{R}$, and $1 \gg \varepsilon > 0, \gamma > 0$ are parameters. In this system, the nonlinear term $f(u)$ is assumed to be a smooth cubic-like function of u satisfying the conditions below.

$$\begin{gathered} f(0) = f(a) = f(1) = 0, \\ f'(0) < 0, \quad f'(1) < 0, \\ f(u) \begin{cases} > 0 & \text{if} \quad u \in (-\infty, 0) \cup (a, 1), \\ < 0 & \text{if} \quad u \in (0, a) \cup (1, +\infty), \end{cases} \\ \textstyle\int_0^1 f(u)du > 0, \end{gathered} \tag{1.2}$$

where $0 < a < 1$ is a constant.

In this paper we shall restrict our attention to large $\gamma > 0$ so that the system (1.1) has three spatially homogeneous stationary solutions $(u, w) \equiv (u_1, w_1) := (0, 0)$, $(u_\dagger, w_\dagger)$, and (u_2, w_2). Here u_* and w_* ($* = 1, 2$ or $\dagger$) are constants which satisfy

$$\begin{gathered} \begin{cases} f(u_*) - w_* &= 0 \\ u_* - \gamma w_* &= 0, \end{cases} \quad * = 1, 2 \text{ or } \dagger \\ 0 = u_1 < u_\dagger < u_2 < 1. \end{gathered} \tag{1.3}$$

(See Figure 1.1.)

The system (1.1) has spatial solutions called travelling waves which are explained below.

*Received by the editors June 28, 1995; accepted for publication May 7, 1996. This research was supported by the JSPS Research Fellowships for Young Scientists and by Science Foundation of the Ministry of Education of Japan grant 3078.

http://www.siam.org/journals/sima/28-5/28829.html

†Department of Mathematics, Faculty of Science, Saitama University, 255 Shimo-Ohkubo, Urawa-shi 338, Japan (snii@rimath.saitama-u.ac.jp).

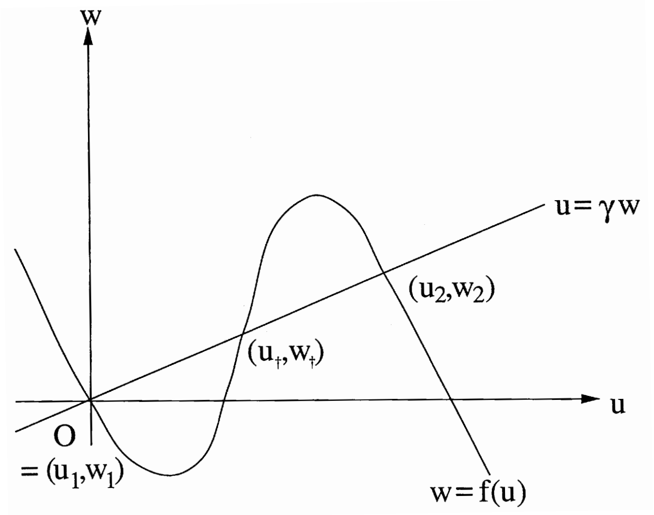


FIG. 1.1. *The nullclines of equations* (1.1).

Let $\xi = x + ct$ be a moving frame for some constant $c > 0$; then in the (ξ, t) coordinate, (1.1) is expressed as

$$\left\{ \begin{array}{ll} u_t & = u_{\xi\xi} - cu_\xi + f(u) - w, \\ w_t & = -cw_\xi + \varepsilon(u - \gamma w). \end{array} \right. \tag{1.4}$$

A travelling wave solution $(u(x,t), w(x,t)) = (u(\xi), w(\xi))$ of (1.1) at velocity c is a steady state solution of (1.4); i.e., $(u(\xi), w(\xi))$ satisfies the equations

$$\left\{ \begin{array}{rl} u_{\xi\xi} - cu_\xi + f(u) - w & = 0, \\ -cw_\xi + \varepsilon(u - \gamma w) & = 0. \end{array} \right. \tag{1.5}$$

Often, (1.5) is treated in the form of first-order equations,

$$\left\{ \begin{array}{ll} u' & = v, \\ v' & = cv - f(u) + w \qquad (' = \frac{d}{d\xi}), \\ w' & = \frac{\varepsilon}{c}(u - \gamma w). \end{array} \right. \tag{1.6}$$

This system shall be simply written as

$$z' = X(z; \mu), \tag{1.7}$$

where $z = (u, v, w)$ and $\mu = (\gamma, c; \varepsilon)$. $a_1 := (u_1, 0, w_1) = (0, 0, 0)$ and $a_2 := (u_2, 0, w_2)$ are equilibria of (1.6).

It is well known that (1.6) has a heteroclinic solution $z_1^*(\xi)$ from a_1 to a_2 ($z_2^*(\xi)$ from a_2 to a_1) for certain parameter values. This solution corresponds to a travelling wave of (1.1), which satisfies

$$\lim_{\xi \to -\infty} z_1^*(\xi) = a_1, \qquad \lim_{\xi \to +\infty} z_1^*(\xi) = a_2 \tag{1.8}$$

$$\left(\lim_{\xi \to -\infty} z_2^*(\xi) = a_2, \lim_{\xi \to +\infty} z_2^*(\xi) = a_1, \text{ respectively} \right). \tag{1.9}$$

This wave is called travelling front, or simple front, in the terminology in Deng [7] (travelling back or simple back, respectively). Deng [7] proved that for certain parameter values, the system (1.6) has heteroclinic solutions z_1^* and z_2^* simultaneously, forming what is called a heteroclinic loop. Furthermore, there is a sequence of N-heteroclinic solutions from a_1 to a_2 (from a_2 to a_1) which correspond to travelling waves called N-fronts (N-backs, respectively) bifurcating from the heteroclinic loop, together with homoclinic solutions to a_1 and a_2 which correspond to travelling pulses (simple impulses, in Deng's terminology). Here a heteroclinic solution from a_1 to a_2 (from a_2 to a_1) which rounds N times and a half in some tubular neighborhood of the heteroclinic loop is referred to as an N-heteroclinic solution from a_1 to a_2 (from a_2 to a_1). (See Figure 1.2.)

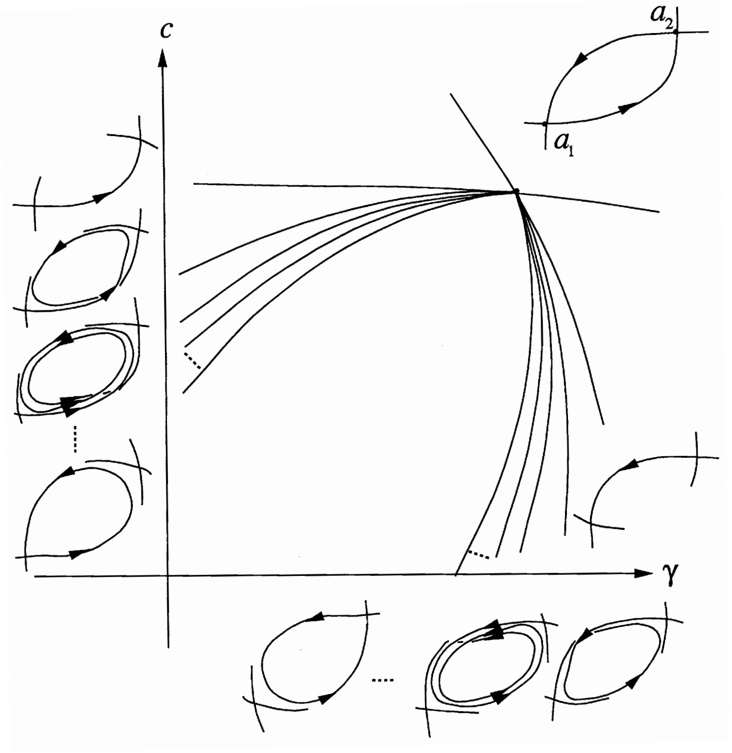


FIG. 1.2. *The bifurcation diagram of N-heteroclinic solutions of the system* (1.6).

We are concerned with the stability of these travelling waves. The eigenvalue problem for (1.4) along the travelling wave under study is often investigated to determine the stability of the wave, as stability for the linear problem implies the same for the full nonlinear problem. See Evans [8].

The linear stability is established as follows. Consider the linearization of (1.4) along the travelling wave $(u(\xi), v(\xi))$ that is under consideration:

$$(1.10) \qquad \begin{cases} P_t &= P_{\xi\xi} - cP_\xi + Df\,(u(\xi))\,P - R, \\ R_t &= -cR_\xi + \varepsilon(P - \gamma R). \end{cases}$$

The right-hand side of (1.10) defines a densely defined closed operator

$$L\begin{pmatrix} P \\ R \end{pmatrix} := \begin{pmatrix} P_{\xi\xi} - cP_\xi + Df\,(u(\xi))\,P - R \\ -cR_\xi + \varepsilon(P - \gamma R) \end{pmatrix} \tag{1.11}$$

on the space $BU(\mathbb{R}, \mathbb{R}^2) := \big\{\phi\colon \mathbb{R} \to \mathbb{R}^2 |\text{ bounded uniformly continuous}\big\}$ with supremum norm. Then, the following fact is well known (Evans [8], Bates and Jones [5]).

Fact. *Let $\sigma(L)$ be the spectrum of L; then the travelling wave $(u(\xi), v(\xi))$ is stable if the conditions below are satisfied.*

1. *There exist $\beta < 0$ so that $\sigma(L) \setminus \{0\} \subset \{\lambda | Re\lambda < \beta\}$.*
2. *0 is a simple eigenvalue.*

Remark 1.1.

1. *L has as an eigenvalue 0 corresponding to spatial translation of the wave.*
2. *Concerning a wave that connects stable steady states, there exists $\beta < 0$ so that $\sigma(L) \cap \{\lambda | \mathrm{Re}\lambda > \beta\}$ consists only of eigenvalues with finite multiplicity.* (*See Jones* [11] *for FitzHugh–Nagumo equations, Henry* [10] *for general cases.*)

Thus we prove the stability of a given wave by showing that zero is a simple eigenvalue of L and that there is no other eigenvalue with zero or positive real part.

The stability of the simple front (back) was proven by Yanagida [17] by showing that the critical eigenvalue was only simple at the origin. The stability of simple impulses can be verified in the same manner as in Jones [11] or Yanagida [16], in which the problem was treated as a singular limit problem for ε when γ was not large, or by applying Nii [12] when it is regarded as a bifurcation problem. In this case the operator L for a simple impulse possesses two critical eigenvalues; one is at the origin, and the sign of the other determines stability. Either way, the Evans function explained below plays a conclusive role.

The eigenvalue problem

$$\begin{cases} P_{\xi\xi} - cP_\xi + Df\,(u(\xi))\,P - R &= \lambda P, \\ \quad -cR_\xi + \varepsilon(P - \gamma R) &= \lambda R \end{cases} \tag{1.12}$$

can be regarded as a system of second-order linear ordinary differential equations. This system shall also be treated in the form of a first-order system,

$$\begin{cases} P' &= Q, \\ Q' &= cQ - Df\,(u(\xi))\,P + \lambda P + R \\ R' &= \frac{\varepsilon}{c}(P - \gamma R) - \frac{\lambda}{c}R, \end{cases} \qquad \left(\ ' = \tfrac{d}{d\xi}\right), \tag{1.13}$$

or simply

$$p' = A\,(u(\xi);\lambda)\,p, \tag{1.14}$$

where $p = (P, Q, R)$ and

$$A\,(u(\xi);\lambda) = \begin{pmatrix} 0 & 1 & 0 \\ \lambda - Df\,(u(\xi)) & c & 1 \\ \frac{\varepsilon}{c} & 0 & -\frac{1}{c}(\varepsilon\gamma + \lambda) \end{pmatrix}.$$

For $\mathrm{Re}\lambda > \beta$ the matrices

$$A_\pm(\lambda) := A(a_{i_\pm};\lambda) = \begin{pmatrix} 0 & 1 & 0 \\ \lambda - Df(a_{i_\pm}) & c & 1 \\ \frac{\varepsilon}{c} & 0 & -\frac{1}{c}(\varepsilon\gamma + \lambda) \end{pmatrix}$$

in both ends ($\xi \to \pm\infty$) of (1.13) have one unstable eigenvalue and two stable ones, where we assume that $\lim_{\xi\to\pm\infty}(u(\xi), w(\xi)) = a_{i_\pm}$. This means (1.13) has one solution $p_1(\xi;\lambda)$ which is bounded as $\xi \to -\infty$ up to multiplication of a nonzero constant and two independent solutions $p_2(\xi;\lambda), p_3(\xi;\lambda)$ which are bounded as $\xi \to +\infty$ up to a nontrivial linear combination of them.

The Evans function $E\mathrm{v}(\lambda)$ is defined as

$$E\mathrm{v}(\lambda) = \det\left(p_1(\xi;\lambda)p_2(\xi;\lambda)p_3(\xi;\lambda)\right)\big|_{\xi=0}. \tag{1.15}$$

Here, as we are working on small ε, $p_1(\xi;\lambda)$, $p_2(\xi,\lambda)$, and $p_3(\xi,\lambda)$ can be chosen so that they depend analytically on λ, and $E\mathrm{v}(\lambda)$ can be defined as an analytic function of λ. By definition, $E\mathrm{v}(\lambda)$ vanishes if and only if $p_1(\xi;\lambda)$, $p_2(\xi;\lambda)$, and $p_3(\xi;\lambda)$ are linearly dependent. This is equivalent to existence of a bounded solution of (1.13), which means that the λ is an eigenvalue of L. If $E\mathrm{v}(\lambda)$ is normalized so that $E\mathrm{v}(\lambda) > 0$ for large $\lambda \in \mathbb{R}$, the sign of $\frac{dE\mathrm{v}}{d\lambda}\big|_{\lambda=0}$ determines the sign of the eigenvalue other than zero. This sign is determined by the geometric structure of the single impulse or corresponding homoclinic orbit. In fact, it is positive and the sign of the eigenvalue is negative. Thus the impulse is stable. In the case of N-front (back) waves, we know that there are $2N+1$ critical eigenvalues near the origin by an argument similar to those in Alexander and Jones [2] or Nii [12]. However, we can only know by an argument similar to that used above that the number of the eigenvalues with positive real part is even. Consequently, we need more analysis to determine their stability.

In this paper, we deal with this eigenvalue problem as a "full" bifurcation problem.

Consider the coupled system of (1.6) and (1.13):

$$\left\{\begin{array}{ll} z' & = X(z;\mu), \\ p' & = A(z;\lambda,\mu)p. \end{array}\right. \tag{1.16}$$

This system on $\mathbb{R}^3 \times \mathbb{C}^3$ induces a system on $\mathbb{R}^3 \times \mathbb{CP}^2$:

$$\left\{\begin{array}{ll} z' & = X(z;\mu), \\ \hat{p}' & = Y(z,\hat{p};\lambda,\mu) \end{array}\right. \tag{1.17}$$

as it is linear in p-component.

Let $e_{i,1}(\lambda)$ ($i = 1, 2$) be an eigenvector associated with the unstable eigenvalue of $A(a_i;\lambda)$ and $e_{i,2}(\lambda)$ and $e_{i,3}(\lambda)$ be eigenvectors associated with the stable eigenvalues. Furthermore, we assume that $e_{i,2}(\lambda)$ belongs to the eigenspace corresponding to the principal stable eigenvalue, which is the stable eigenvalue with its real part larger than the other. The points in $\mathbb{CP}^2$ representing eigenspaces spanned by $e_{i,j}(\lambda)$ shall be denoted as $\hat{e}_{i,j}(\lambda)$. Then for each $i = 1, 2$, $\{a_i\} \times \mathbb{CP}^2$ is an invariant set of (1.17), which consists of equilibria $(a_i, \hat{e}_{i,j}(\lambda))$ ($j = 1, 2, 3$) and heteroclinic orbits between them. For the parameter value μ at which (1.6) has a heteroclinic solution from a_{i_-} to a_{i_+}, the system (1.17) should have a heteroclinic solution from $\left(a_{i_-}, \hat{e}_{i_-,1}(\lambda)\right)$ to $\left(a_{i_+}, \hat{e}_{i_+,j}(\lambda)\right)$ for some j depending on λ. For generic λ this solution should be from $\left(a_{i_-}, \hat{e}_{i_-,1}(\lambda)\right)$ to $\left(a_{i_+}, \hat{e}_{i_+,1}(\lambda)\right)$, because $\left(a_{i_+}, \hat{e}_{i_+,1}(\lambda)\right)$ is an attracting equilibrium in the invariant set $\{a_{i_+}\} \times \mathbb{CP}^2$ and the complementary repeller consists of $\left(a_{i_+}, \hat{e}_{i_+,2}(\lambda)\right)$, $\left(a_{i_+}, \hat{e}_{i_+,3}(\lambda)\right)$, and the heteroclinic orbits from $\left(a_{i_+}, \hat{e}_{i_+,3}(\lambda)\right)$ to $\left(a_{i_+}, \hat{e}_{i_+,3}(\lambda)\right)$. In fact, the existence of the solution from $\left(a_{i_-}, \hat{e}_{i_-,1}(\lambda)\right)$ to $\left(a_{i_+}, \hat{e}_{i_+,2}(\lambda)\right)$ or $\left(a_{i_+}, \hat{e}_{i_+,3}(\lambda)\right)$ means that λ is an eigenvalue of L and vice versa.

Let μ_0 be a parameter value at which (1.6) has a heteroclinic loop consisting of heteroclinic solutions $z_1^*(\xi)$ from a_1 to a_2 and $z_2^*(\xi)$ from a_2 to a_1. Then, for

$(\lambda, \mu) = (0, \mu_0)$, (1.17) has heteroclinic solutions from $(a_1, \hat{e}_{1,1}(0))$ to $(a_2, \hat{e}_{2,2}(0))$ and from $(a_2, \hat{e}_{2,1}(0))$ to $(a_1, \hat{e}_{1,2}(0))$ simultaneously. We interpret the eigenvalue problem associated with an N-front wave which corresponds to an N-heteroclinic solution from a_{i_-} to a_{i_+} as a bifurcation problem of finding an N-heteroclinic solution of (1.17) from $\big(a_{i_-}, \hat{e}_{i_-,1}(\lambda)\big)$ to $\big(a_{i_+}, \hat{e}_{i_+,2}(\lambda)\big)$ or $\big(a_{i_+}, \hat{e}_{i_+,3}(\lambda)\big)$.

The purpose of this paper is to prove the following theorem about the stability of the travelling 1-front (back) wave solutions of (1.1), with the strategy explained above. The stability of N-fronts for $N \geq 2$ shall be proven in the forthcoming paper using a topological method (Nii [13]).

THEOREM. *Assume that the system* (1.6) *is C^r-diffeomorphic* ($r \geq 2$) *to linear systems in some neighborhoods of equilibria a_i and ε is small; then the travelling* 1*-front* (*back*) *wave solution of FitzHugh–Nagumo equations* (1.1) *bifurcating from simple front and back travelling wave solutions is stable.*

REMARK 1.2. *The assumption that the system is diffeomorphic to linear systems shall also be discussed for strictly cubic nonlinearity.*

REMARK 1.3. *Recently, Sandstede* [15] *proved stability of N-fronts* (*backs*) *for all $N \geq 1$ by Lin's method.*

2. Existence of travelling N-front wave solutions. In this section we briefly summarize the result in Deng [7] concerning the existence of travelling N-front (back) wave solutions of (1.1) and give analysis of the parameter dependence of the 1-heteroclinic orbit of (1.6) around equilibria.

PROPOSITION 2.1 (Deng [7]). *There exists a small ε_0 and two smooth functions $\gamma(\varepsilon)$ and $\delta(\varepsilon)$ for $0 \leq \varepsilon \leq \varepsilon_0$ such that the following is satisfied for all $0 < \varepsilon < \varepsilon_0$.*

1. *On the relevant (γ, c) parameter space there are two smooth curves $c = c_{i,0}(\gamma)$* ($i = 1, 2$) *defined on the interval $|\gamma - \gamma(\varepsilon)| < \delta(\varepsilon)$ such that* (1.1) *has a simple front wave of speed $c_{1,0}(\gamma)$ and a simple back wave of speed $c_{2,0}(\gamma)$.*
2. *There is a sequence $\{c_{1,N}(\gamma)\}_{N=1}^{\infty}$ of smooth curves of the left half-interval $0 < \gamma(\varepsilon) - \gamma < \delta(\varepsilon)$ such that* (1.1) *has an N-front wave of speed $c_{1,N}(\gamma)$ for γ in this half-interval for every $N = 1, 2, \ldots$. Similarly, there is a sequence $\{c_{2,N}(\gamma)\}_{N=1}^{\infty}$ of smooth curves of the right half-interval $0 < \gamma - \gamma(\varepsilon) < \delta(\varepsilon)$ such that* (1.1) *has an N-back wave of speed $c_{2,N}(\gamma)$ for γ in this half-interval for every $N = 1, 2, \ldots$.*
3. *There is a smooth curve $c_{1,\infty}(\gamma)$ of the left half-interval $0 < \gamma(\varepsilon) - \gamma < \delta(\varepsilon)$ such that* (1.1) *has an impulse wave to a_1 with speed $c_{1,\infty}(\gamma)$. Similarly, there is a smooth curve $c_{2,\infty}(\gamma)$ of the right half-interval $0 < \gamma - \gamma(\varepsilon) < \delta(\varepsilon)$ such that* (1.1) *has an impulse wave to a_2 with speed $c_{2,\infty}(\gamma)$.*
4. *The simple front and back wave curves $c_{i,0}(\gamma)$ intersect transversely at $\gamma(\varepsilon)$. The intersection point $(\gamma(\varepsilon), c(\varepsilon))$ is smooth in ε. At $\varepsilon = 0$, $c(0) > 0$. Moreover, the sequence $\{c_{i,N}(\gamma)\}$ is monotone decreasing in $N = 1, 2, \ldots$ and converges to the corresponding impulse curve $c_{i,\infty}(\gamma)$ as $N \to \infty$ for each $i = 1, 2$ and fixed $\varepsilon > 0$ and γ.*

(*See Figure* 2.1.)

REMARK 2.1. *This result is proven in Deng* [7] *for cubic nonlinearity $f(u) = u(u-a)(1-u)$ for* (1.1)*; the proof given there, however, is valid for $f(u)$, which is not necessarily cubic. The condition $0 < a < \frac{1}{2}$ in* [7] *corresponds to the condition $\int_0^1 f(u)du > 0$ in this paper.*

We need more detailed information about the heteroclinic orbit of (1.6) corresponding to the 1-front wave around equilibria.

In what follows, we assume that the system (1.6) is linear in some small neigh-

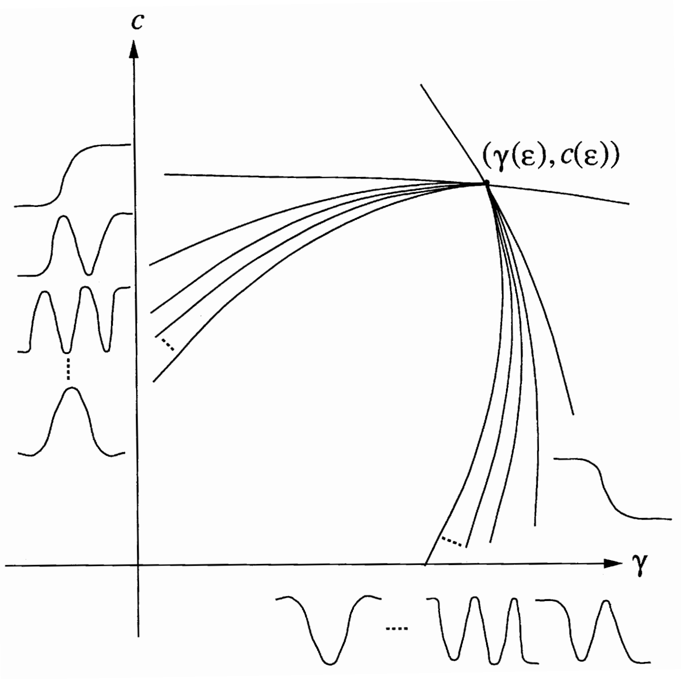


FIG. 2.1. *The bifurcation diagram of the travelling* N*-front (back) wave solutions.*

borhood of equilibria a_i $(i = 1, 2)$. This assumption can be satisfied for strictly cubic nonlinearity for suitable parameter values. In fact, the following holds.

PROPOSITION 2.2. *Let* $f(u) = u(u - a)(1 - u)$*; then there are uncountably many pairs of* (a, ε) $(0 < a < \frac{1}{2}, 0 < \varepsilon < \varepsilon_0)$ *such that for all* r *there is a neighborhood* N *of* $(\gamma(\varepsilon), c(\varepsilon))$ *in the* γ*–*c *plane depending on* a, ε, *and* r *which has the following property. There are small neighborhoods of equilibria* a_i *in which* (1.6) *can be transformed into a linear system by a local* C^r*-coordinate change in the neighborhoods.*

The proof shall be given in the Appendix.

Through suitable local coordinate changes, let us assume that in some neighborhood of each equilibrium a_i the system (1.6) is expressed in the form

$$z_i' = D_i(\mu)z_i, \tag{2.1}$$

where $z_i = (z_i^{(1)}, z_i^{(2)}, z_i^{(3)})$ and

$$D_i(\mu) = \begin{pmatrix} \lambda_{i,u}(\mu) & 0 & 0 \\ 0 & -\lambda_{i,s}(\mu) & 0 \\ 0 & 0 & -\lambda_{i,ss}(\mu) \end{pmatrix},$$

with $-\lambda_{i,ss}(\mu) < -\lambda_{i,s}(\mu) < 0 < \lambda_{i,u}(\mu)$. Moreover, the cube $\max_{k=1,2,3} |z_i^{(k)}| \leq 1$ is assumed to be in each neighborhood, and the orientation of axes is chosen so that $\{z_i | z_i^{(1)} > 0, z_i^{(2)} = 0, z_i^{(3)} = 0\} \subset \{z_i^*(\xi) | \xi \in \mathbb{R}\}$ and $\{z_i | z_i^{(1)} = 0, z_i^{(2)} > 0, z_i^{(3)} = 0\} \subset \{z_j^*(\xi) | \xi \in \mathbb{R}\}$ $(i \neq j)$ holds for $(\gamma, c) = (\gamma(\varepsilon), c(\varepsilon))$.

With this coordinate in z and through a suitable analytic transformation in the p component, (1.16) is written as

$$\left\{ \begin{aligned} z_i' &= D_i(\mu)z_i, \\ p_i' &= A_i(\lambda, \mu)p_i \end{aligned} \right. \tag{2.2}$$

for small $|\lambda|$, where $p_i = (p_i^{(1)}, p_i^{(2)}, p_i^{(3)})$ and

$$A_i(\lambda,\mu) = \begin{pmatrix} \nu_{i,u}(\lambda,\mu) & 0 & 0 \\ 0 & -\nu_{i,s}(\lambda,\mu) & 0 \\ 0 & 0 & -\nu_{i,ss}(\lambda,\mu) \end{pmatrix},$$

with $\nu_{i,*}(0,\mu) = \lambda_{i,*}(\mu)$ $(* = u, s, ss)$. The projectivized version of this system shall be denoted by

$$\tag{2.3} \begin{cases} z_i' &= D_i(\mu) z_i, \\ \hat{p}_i' &= \hat{A}_i(\hat{p}_i; \lambda, \mu). \end{cases}$$

A solution of (2.2) with initial condition $(z_i, p_i) = (\delta, 1, \bar{\delta}, p^{(1)}, p^{(2)}, p^{(3)})$ $(0 < \delta, \bar{\delta} < 1)$ hits $(1, \delta^{\Lambda_{i,s}}, \bar{\delta}\delta^{\Lambda_{i,ss}}, p^{(1)}\delta^{-(\nu_{i,u}/\lambda_{i,u})}, p^{(2)}\delta^{\nu_{i,s}/\lambda_{i,u}}, p^{(3)}\delta^{\nu_{i,ss}/\lambda_{i,u}})$, where $\Lambda_{i,s} = \frac{\lambda_{i,s}}{\lambda_{i,u}}$ and $\Lambda_{i,ss} = \frac{\lambda_{i,ss}}{\lambda_{i,u}}$. As for (2.3), we employ inhomogeneous coordinates on $\mathbb{CP}^2$. More precisely, the initial point $(z_i, \hat{p}_i) = (\delta, 1, \bar{\delta}, [p_i^{(1)} : p_i^{(2)} : p^{(3)}])$ in homogeneous coordinates shall be expressed as $(\delta, 1, \bar{\delta}, \frac{p_i^{(1)}}{p_i^{(2)}}, \frac{p_i^{(3)}}{p_i^{(2)}})$, whereas the end point $(1, \delta^{\Lambda_{i,s}}, \bar{\delta}\delta^{\Lambda_{i,ss}}, [p_i^{(1)}\delta^{-(\nu_{i,u}/\lambda_{i,u})} : p_i^{(2)}\delta^{\nu_{i,s}/\lambda_{i,u}} : p_i^{(3)}\delta^{\nu_{i,ss}/\lambda_{i,u}}])$ shall be denoted by $(1, \delta^{\Lambda_{i,s}}, \bar{\delta}\delta^{\Lambda_{i,ss}}, \frac{p^{(2)}}{p^{(1)}}\delta^{1+\hat{\Lambda}_{i,s}}, \frac{p^{(3)}}{p^{(1)}}\delta^{1+\hat{\Lambda}_{i,ss}})$. Notice that the two expressions above are based on different coordinates for the $\mathbb{CP}^2$ component. Here, $\hat{\Lambda}_{i,s} = \frac{\nu_{i,s}+\nu_{i,u}-\lambda_{i,u}}{\lambda_{i,u}}$ and $\hat{\Lambda}_{i,ss} = \frac{\nu_{i,ss}+\nu_{i,u}-\lambda_{i,u}}{\lambda_{i,u}}$, which coincide with $\Lambda_{i,s}$ and $\Lambda_{i,ss}$ for $\lambda = 0$, and thus $\Lambda_{i,s} < \Lambda_{i,ss}$ and $0 < \Lambda_{i,s} < 1$ for small λ for small ε. We restate this calculation as a lemma.

LEMMA 2.1. *The solution of* (2.3) *with initial condition* $(z, \frac{p_i^{(1)}}{p_i^{(2)}}, \frac{p_i^{(3)}}{p_i^{(2)}}) = (\delta, 1, \bar{\delta}, \pi_1, \pi_3)$ *reaches* $(z, \frac{p_i^{(2)}}{p_i^{(1)}}, \frac{p_i^{(3)}}{p_i^{(1)}}) = (1, \delta^{\Lambda_{i,s}}, \bar{\delta}\delta^{\Lambda_{i,ss}}, \frac{1}{\pi_1}\delta^{1+\hat{\Lambda}_{i,s}}, \frac{\pi_3}{\pi_1}\delta^{1+\hat{\Lambda}_{i,ss}})$.

Next, we consider connections between a_1 and a_2.

Let $\Sigma_{i,s}$ and $\Sigma_{i,u}$ be the local sections near a_i defined as

$$\tag{2.4} \begin{aligned} \Sigma_{i,s} &:= \left\{ z_i^{(2)} = 1, \max\{|z_i^{(1)}|, |z_i^{(3)}|\} \leq 1 \right\}, \\ \Sigma_{i,u} &:= \left\{ z_i^{(1)} = 1, \max\{|z_i^{(2)}|, |z_i^{(3)}|\} \leq 1 \right\}, \end{aligned}$$

and let

$$\tag{2.5} \begin{aligned} \Pi_i &: \Sigma_{i,s} \to \Sigma_{i,u} : (z_i^{(2)}, z_i^{(3)}) = \left(\Pi_i^{(2)}(z_i^{(1)}, z_i^{(3)}), \Pi_i^{(3)}(z_i^{(1)}, z_i^{(3)}) \right), \\ \Pi_{i,j} &: \Sigma_{i,u} \to \Sigma_{j,s} : (z_j^{(1)}, z_j^{(3)}) = \left(\Pi_{i,j}^{(1)}(z_i^{(2)}, z_i^{(3)}), \Pi_{i,j}^{(3)}(z_i^{(2)}, z_i^{(3)}) \right) \end{aligned}$$

be the Poincaré maps between them, where $(i,j) = (1,2)$ or $(2,1)$. (See Figure 2.2.)

By Proposition 2.1, the system (1.6) has heteroclinic solutions $z_1^*(\xi)$ from a_1 to a_2 and $z_2^*(\xi)$ from a_2 to a_1 for $\mu = (\gamma(\varepsilon), c(\varepsilon); \varepsilon)$, which means

$$\tag{2.6} \Pi_{i,j}^{(1)}(0, 0; \gamma(\varepsilon), c(\varepsilon); \varepsilon) = 0 \qquad \text{for } (i,j) = (1,2), (2,1).$$

The twisting condition of the heteroclinic loop in Deng [6], [7] is equivalent to the inequality

$$\tag{2.7} \frac{\partial \Pi_{i,j}^{(1)}}{\partial z_i^{(2)}}(0, 0; \gamma(\varepsilon), c(\varepsilon); \varepsilon) < 0.$$

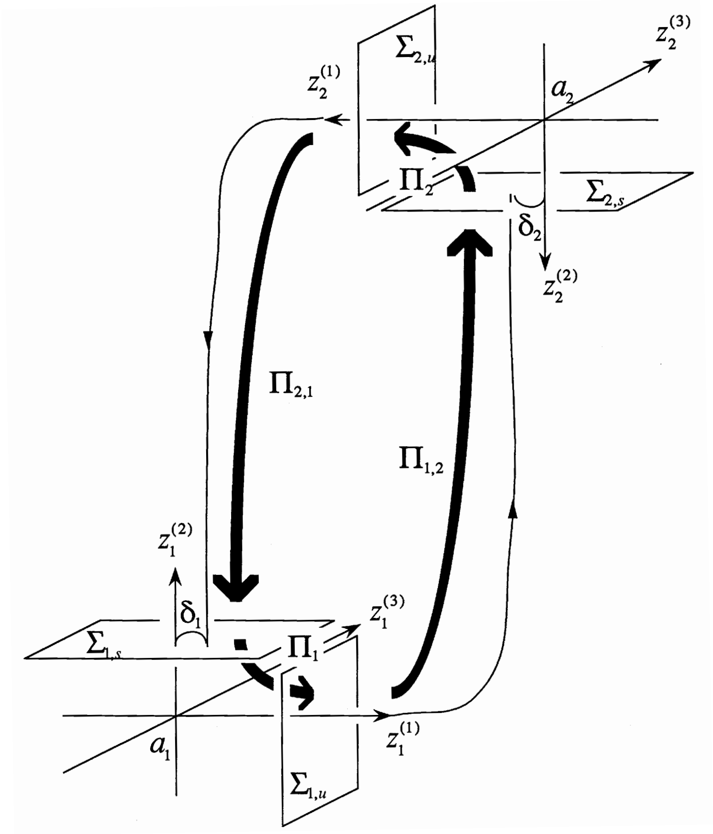


FIG. 2.2. *The local sections and Poincaré maps.*

The following transversality condition was also proven in Deng [7]:

$$\left(\frac{\partial \Pi^{(1)}_{i,j}}{\partial \gamma}(0,0), \frac{\partial \Pi^{(1)}_{i,j}}{\partial c}(0,0)\right) \neq (0,0) \qquad ((i,j) = (1,2) \text{ or } (2,1), \quad \varepsilon > 0)\,. \tag{2.8}$$

Moreover, $(\frac{\partial \Pi^{(1)}_{1,2}}{\partial \gamma}(0,0), \frac{\partial \Pi^{(1)}_{1,2}}{\partial c}(0,0))$ and $(\frac{\partial \Pi^{(1)}_{2,1}}{\partial \gamma}(0,0), \frac{\partial \Pi^{(1)}_{2,1}}{\partial c}(0,0))$ are linearly independent, and thus we can employ $(\Pi^{(1)}_{1,2}(0,0), \Pi^{(1)}_{2,1}(0,0))$ as parameters instead of (γ, c) near $(\gamma, c) = (\gamma(\varepsilon), c(\varepsilon))$. In what follows, we shall write $(\Pi^{(1)}_{2,1}(0,0), \Pi^{(1)}_{1,2}(0,0)) = (\delta_1, \delta_2)$ and regard $\varepsilon > 0$ as a fixed constant. Then, for z_i and δ_i small, Π_i and $\Pi_{i,j}$ are written as

$$\left(\Pi_i^{(2)}(z_i^{(1)}, z_i^{(3)}), \Pi_i^{(3)}(z_i^{(1)}, z_i^{(3)})\right) = \left(\left\{z_i^{(1)}\right\}^{\Lambda_{i,s}}, z_i^{(3)}\left\{z_i^{(1)}\right\}^{\Lambda_{i,ss}}\right) \tag{2.9}$$

and

$$\begin{pmatrix} \Pi^{(1)}_{i,j}(z_i^{(2)}, z_i^{(3)}; \delta_1, \delta_2) \\ \Pi^{(3)}_{i,j}(z_i^{(2)}, z_i^{(3)}; \delta_1, \delta_2) \end{pmatrix} = \begin{pmatrix} d_j^{(1)} z_i^{(2)} + \bar{d}_j^{(1)} z_i^{(3)} + \delta_j \\ d_j^{(3)} z_i^{(2)} + \bar{d}_j^{(3)} z_i^{(3)} + e_j^1 \delta_1 + e_j^2 \delta_2 \end{pmatrix} \tag{2.10}$$

$$+(\text{higher order terms}),$$

where $d_j^{(k)} = \frac{\partial \Pi_{i,j}^{(k)}}{\partial z_i^{(2)}}(0,0;0,0)$, $\bar{d}_j^{(k)} = \frac{\partial \Pi_{i,j}^{(k)}}{\partial z_i^{(3)}}(0,0;0,0)$, and $e_j^l = \frac{\partial \Pi_{i,j}^{(3)}}{\partial \delta_l}(0,0;0,0)$.

The existence of a 1-heteroclinic solution from a_1 to a_2 is expressed as

$$\begin{gathered} \Pi_{1,2}^{(1)}(0,0) > 0,\ \Pi_{2,1}^{(1)} \circ \Pi_2 \circ \Pi_{1,2}(0,0) > 0, \\ \Pi_{1,2}^{(1)} \circ \Pi_1 \circ \Pi_{2,1} \circ \Pi_2 \circ \Pi_{1,2}(0,0) = 0. \end{gathered} \tag{2.11}$$

Put $(\delta_1^{(1)}, \delta_1^{(3)}) := \Pi_{2,1} \circ \Pi_2 \circ \Pi_{1,2}(0,0)$; then

$$\Pi_1(\delta_1^{(1)}, \delta_1^{(3)}) = \left(\left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}}, \delta_1^{(3)} \left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,ss}} \right) \tag{2.12}$$

and

$$\begin{aligned} \Pi_{1,2}^{(1)} \circ \Pi_1 \left(\delta_1^{(1)}, \delta_1^{(3)} \right) = \ & \left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}} \left(d_2^{(1)} + \bar{d}_2^{(1)} \delta_1^{(3)} \left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,ss} - \Lambda_{1,s}} \right) \\ & + o \left(\left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}} \right) + \delta_2 + o\left(\delta_2\right), \end{aligned} \tag{2.13}$$

so the last equation of (2.11) is

$$\begin{aligned} \left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}} & \left(d_2^{(1)} + \bar{d}_2^{(1)} \delta_1^{(3)} \left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,ss} - \Lambda_{1,s}} \right) \\ & + o \left(\left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}} \right) + \delta_2 + o\left(\delta_2\right) = 0. \end{aligned} \tag{2.14}$$

By (2.10) we have the following expressions:

$$\begin{aligned} \delta_1^{(1)} & = d_1^{(1)} \delta_2^{\Lambda_{2,s}} + \bar{d}_1^{(1)} \left(e_2^1 \delta_1 + e_2^2 \delta_2 \right) \delta_2^{\Lambda_{2,ss}} + \delta_1 \\ & \qquad + \text{higher order terms}, \\ \delta_1^{(3)} & = d_1^{(3)} \delta_2^{\Lambda_{2,s}} + \bar{d}_1^{(3)} \left(e_2^1 \delta_1 + e_2^2 \delta_2 \right) \delta_2^{\Lambda_{2,ss}} + e_1^1 \delta_1 + e_1^2 \delta_2 \\ & \qquad + \text{higher order terms}. \end{aligned} \tag{2.15}$$

Here, the bifurcation curve $\{(\gamma, c_{1,1}(\gamma))\}$ is expressed as $\delta_1 = \mathrm{het}_1(\delta_2)$ with a smooth function $\mathrm{het}_1\colon (0, \Delta_2) \to (0, \Delta_1)$ in the (δ_1, δ_2)-coordinates for some $(\Delta_1, \Delta_2 > 0)$, and this function satisfies

$$\lim_{\delta_2 \downarrow 0} \mathrm{het}_1(\delta_2) = 0. \tag{2.16}$$

Therefore, $\delta_1^{(1)}, \delta_2^{(1)} \to 0$ as $\delta_2 \to 0$. Then, (2.14) means

$$\frac{\left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}}}{\delta_2} (1 + o(1)) + \frac{1}{d_2^{(1)}} = o(1) \qquad (\delta_2 \to 0), \tag{2.17}$$

i.e.,

$$\frac{\left\{ \delta_1^{(1)} \right\}^{\Lambda_{1,s}}}{\delta_2} \to \frac{1}{-d_2^{(1)}} \quad \text{as} \quad \delta_2 \to 0. \tag{2.18}$$

Thus the following holds.

LEMMA 2.2.

$$\delta_1^{(1)} = \left(\frac{\delta_2}{-d_2^{(1)}} \right)^{\frac{1}{\Lambda_{1,s}}} + o\left(\delta_2^{\frac{1}{\Lambda_{1,s}}} \right). \tag{2.19}$$

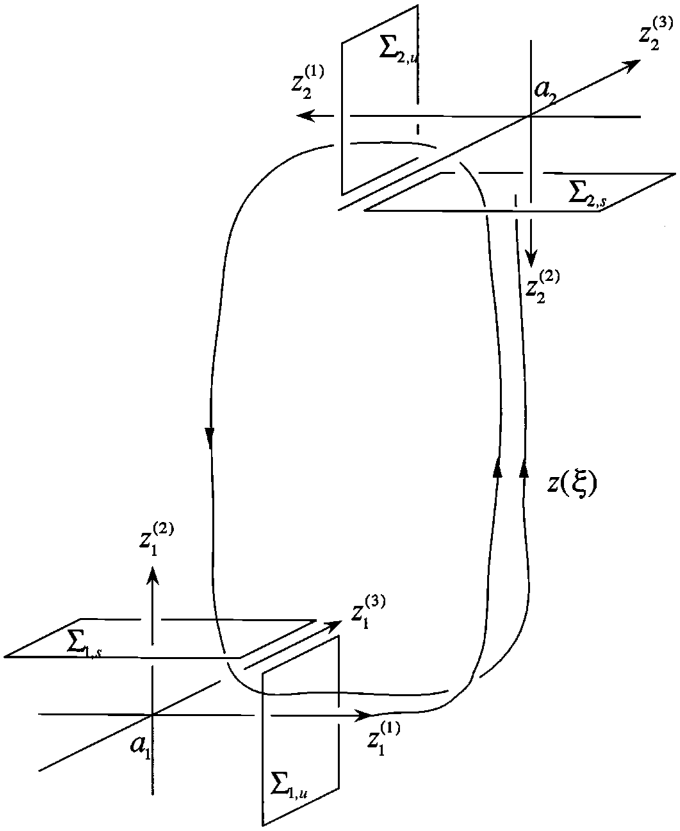


FIG. 3.1. *The orbit of* $z(\xi;\gamma,c)$.

3. λ-dependence of the eigenvalue system.

In this section, we investigate λ-dependence of the eigenvalue system (1.17).

Let us assume that the system (1.6) has a heteroclinic orbit from a_i to a_j $(i \neq j)$ for $(\gamma, c) = (\gamma_0, c_0)$, and let $z_i(\xi;\gamma_0,c_0) = (u(\xi;\gamma_0,c_0), v(\xi;\gamma_0,c_0), w(\xi;\gamma_0,c_0))$ be the solution with initial condition $z_i(0;\gamma_0,c_0) \in \Sigma_{j,s}$, or equivalently, $\{z_i(0;\gamma_0,c_0)\} = \Sigma_{j,s} \cap W^u(a_i,\gamma_0,c_0)$, where $W^u(a_i,\gamma_0,c_0)$ is the unstable manifold of the equilibrium a_i. Similarly, let $z(\xi;\gamma,c)$ be the solution of (1.6) with initial condition $\{z(0;\gamma,c)\} = \Sigma_{j,s} \cap W^u(a_i,\gamma,c)$ for (γ,c) near (γ_0,c_0); then $z(\xi;\gamma,c)$ is smooth in γ and c and coincides with $z_i(\xi;\gamma_0,c_0)$ for $(\gamma,c) = (\gamma_0,c_0)$. (See Figure 3.1.)

By differentiating (1.6) in c at $(\gamma,c) = (\gamma_0,c_0)$, $\frac{\partial z_i}{\partial c} =: z_{ic} = (u_c, v_c, w_c)$ satisfies

$$\left\{\begin{array}{ll} u_c' & = v_c, \\ v_c' & = v - c_0 v_c - Df\,(u(\xi))\, u_c + w_c, \\ w_c' & = \frac{\varepsilon}{c_0}(u_c - \gamma_0 w_c) - \frac{1}{c_0} w', \end{array}\right. \tag{3.1}$$

where $(u,v,w) = (u(\xi;\gamma_0,c_0), v(\xi;\gamma_0,c_0), w(\xi;\gamma_0,c_0))$ and $w' = \frac{dw}{d\xi}$.

Next consider the restriction of (1.16) to $\mathbb{R}^3 \times \mathbb{R}^3$ for $(\gamma,c) = (\gamma_0,c_0)$. The equilibrium $(a_i,0)$ of (1.16) has two-dimensional unstable manifold $\tilde{W}^u(a_i,\lambda,\gamma_0,c_0)$, and $\tilde{W}^u(a_i,\lambda,\gamma_0,c_0) \cap \left(\{z_i(0,\gamma_0,c_0)\} \times \mathbb{R}^3\right)$ is a one-dimensional subspace of $\mathbb{R}^3$. For $\lambda = 0$, $(z(0;\gamma_0,c_0), z_\xi(0;\gamma_0,c_0)) \in \tilde{W}^u(a_i,0,\gamma_0,c_0) \cap \left(\{z(0,\gamma_0,c_0)\} \times \mathbb{R}^3\right)$, where $z_\xi =$

$\frac{dz}{d\xi}$. Let $(z(\xi;\gamma_0,c_0),p_i(\xi;\lambda,\gamma_0,c_0))$ be the solution of (1.16) with initial condition

$$(z(0),p(0)) = \tilde{W}^u(a_i,\lambda,\gamma_0,c_0) \cap \left(\{z(0,\gamma_0,c_0)\} \times \mathbb{R}^3\right) \cap \{|p| = |z_\xi(0)|\}\,;$$

then $p_i(\xi;\lambda,\gamma_0,c_0)$ coincides with $z_\xi(\xi;\gamma_0,c_0)$ for $\lambda = 0$ and is smooth in small λ. This $p_i(\xi;\lambda,\gamma_0,c_0)$ satisfies (1.13), and by differentiating (1.13) with respect to λ at $\lambda = 0$, $\frac{\partial p_i}{\partial \lambda} =: p_{i\lambda} = (P_\lambda, Q_\lambda, R_\lambda)$ satisfies the following equations:

$$\text{(3.2)} \qquad \left\{ \begin{array}{ll} P'_\lambda & = Q_\lambda, \\ Q'_\lambda & = c_0 Q_\lambda - Df\left(u(\xi)\right) P_\lambda + R_\lambda + P, \\ R'_\lambda & = \frac{\varepsilon}{c_0}(P_\lambda - \gamma_0 R_\lambda) - \frac{1}{c_0} R', \end{array} \right.$$

where $(P,Q,R) = p_i(\xi;0,\gamma_0,c_0) = z_\xi(\xi;\gamma_0,c_0) = (u',v',w')$. Subtracting (3.1) from (3.2), we have

$$\text{(3.3)} \qquad \left\{ \begin{array}{ll} (P_\lambda - u_c)' & = (Q_\lambda - v_c), \\ (Q_\lambda - v_c)' & = c_0(Q_\lambda - v_c) - Df\left(u(\xi)\right)(P_\lambda - u_c) + (R_\lambda - w_c), \\ (R_\lambda - w_c)' & = \frac{\varepsilon}{c_0}\left((P_\lambda - u_c) - \gamma_0(R_\lambda - w_c)\right), \end{array} \right.$$

as $P = u' = v$ and $R = w'$. That is, $p_{i\lambda}(\xi;0,\gamma_0,c_0) - z_c(\xi;\gamma_0,c_0)$ satisfies the eigenvalue equations (1.13) for $\lambda = 0$. Because $p_{i\lambda}, z_c \to 0$ as $\xi \to -\infty$, we have the next lemma.

LEMMA 3.1. *There exists a constant α so that*

$$p_{i\lambda}(\xi;0,\gamma_0,c_0) = z_c(\xi;\gamma_0,c_0) + \alpha z_{i\xi}(\xi;\gamma_0,c_0).$$

Similarly, let $(z_i(\xi;\gamma_0,c_0),q_i(\xi;\lambda,\gamma_0,c_0))$ be the solution of (2.2) (see (1.16)) with initial condition

$$(z(0),q(0)) = \tilde{W}^s(a_j,\lambda,\gamma_0,c_0) \cap \left(\{z_i(0,\gamma_0,c_0)\} \times \mathbb{R}^3\right) \cap \left\{p_j^{(2)} = z_{i\xi}^{(2)}(0), p_j^{(3)} = z_{i\xi}^{(3)}(0)\right\},$$

where $\tilde{W}^s(a_j,\lambda,\gamma_0,c_0)$ is the stable manifold of the equilibrium $(a_j,0)$ of (2.2) and the p_i component of (2.2) is written as $p_i = (p_i^{(1)},p_i^{(2)},p_i^{(3)}) \in \mathbb{R}^3$. Then $q_i(\xi;\lambda,\gamma_0,c_0)$ coincides with $z_{i\xi}(\xi;\gamma_0,c_0)$ for $\lambda = 0$ and is smooth in small λ. Through a similar argument as above, we have the following as $z(\xi;\gamma,c)$ with initial condition $z(0;\gamma,c) \equiv z_i(0;\gamma_0,c_0)$ in the z_j coordinate of (2.2) is a smooth extension of $z_i(\xi;\gamma_0,c_0)$ in the stable manifold $W^s(a_j,\gamma,c)$.

LEMMA 3.2.

$$(z_i(\xi;\gamma_0,c_0),q_{i\lambda}(\xi;0,\gamma_0,c_0)) \in \tilde{W}^s(a_j,0,\gamma_0,c_0).$$

From now on, we consider the projectivized version of eigenvalue systems (1.17) and (2.3).

At first we only consider real λ and consider the system (1.17) as the system on $\mathbb{R}^3 \times \mathbb{RP}^2$. Let $\hat{\Sigma}_{i,s}$ and $\hat{\Sigma}_{i,u}$ be local sections defined by $\hat{\Sigma}_{i,s} := \Sigma_{i,s} \times \mathbb{RP}^2$, $\hat{\Sigma}_{i,u} := \Sigma_{i,u} \times \mathbb{RP}^2$, and let

$$\text{(3.4)} \qquad \begin{array}{ll} \hat{\Pi}_i & : \hat{\Sigma}_{i,s} \to \hat{\Sigma}_{i,u} \\ & : (z_i^{(2)}, z_i^{(3)}, \frac{p_i^{(2)}}{p_i^{(1)}}, \frac{p_i^{(3)}}{p_i^{(1)}}) = \left(\Pi_i^{(2)}(z_i), \Pi_i^{(3)}(z_i), \hat{\Pi}_i^{(2)}(z_i,p_i), \hat{\Pi}_i^{(3)}(z_i,p_i)\right), \\ \hat{\Pi}_{i,j} & : \hat{\Sigma}_{i,u} \to \hat{\Sigma}_{j,s} \\ & : (z_j^{(1)}, z_j^{(3)}, \frac{p_j^{(1)}}{p_j^{(2)}}, \frac{p_j^{(3)}}{p_j^{(2)}}) = \left(\Pi_{i,j}^{(1)}(z_i), \Pi_{i,j}^{(3)}(z_i), \hat{\Pi}_{i,j}^{(1)}(z_i,p_i), \hat{\Pi}_{i,j}^{(3)}(z_i,p_i)\right) \end{array}$$

be Poincaré maps between them. Then, from Lemma 2.1,

$$\begin{aligned}&\hat{\Pi}_i(z_i^{(1)}, z_i^{(3)}, \pi_1, \pi_3)\\ &= \left(\left\{z_i^{(1)}\right\}^{\Lambda_{i,s}}, z_i^{(3)}\left\{z_i(1)\right\}^{\Lambda_{i,ss}}, \tfrac{1}{\pi_1}\left\{z_i^{(1)}\right\}^{1+\hat{\Lambda}_{i,s}}, \tfrac{\pi_3}{\pi_1}\left\{z_i^{(1)}\right\}^{1+\hat{\Lambda}_{i,ss}}\right).\end{aligned} \tag{3.5}$$

As for $\hat{\Pi}_{i,j}$, the following holds.

Lemma 3.3.

$$\hat{\Pi}_{i,j}^{(1)}(0,0;0) = 0, \quad \frac{\partial \hat{\Pi}_{i,j}^{(1)}}{\partial \pi_2}(0,0;0) > 0, \quad \textit{and} \quad \frac{\partial \hat{\Pi}_{i,j}^{(1)}}{\partial \lambda}(0,0;0) > 0, \tag{3.6}$$

where $(z_i, \pi; \lambda) = (z_i^{(2)}, z_i^{(3)}, \pi_2 = \frac{p_i^{(2)}}{p_i^{(1)}}, \pi_3 = \frac{p_i^{(3)}}{p_i^{(1)}}; \lambda)$.

Proof. The first equality is obvious from the definition. As for the second inequality, it is easy to see that

$$\frac{\partial \hat{\Pi}_{i,j}^{(1)}}{\partial \pi_2}(0,0;0) = -\frac{\lambda_{i,u}}{\lambda_{j,s}} \frac{\partial \Pi_{i,j}^{(1)}}{\partial z_i^{(2)}}(0,0);$$

thus the inequality follows from inequality (2.7).

A proof of the last inequality shall be given below.

Let

$$p_i(\xi, \lambda, \gamma_0, c_0) = \left(p_i^{(1)}(\xi,\lambda), p_i^{(2)}(\xi,\lambda), p_i^{(3)}(\xi,\lambda)\right)$$

and

$$q_i(\xi, \lambda, \gamma_0, c_0) = \left(q_i^{(1)}(\xi,\lambda), q_i^{(2)}(\xi,\lambda), q_i^{(3)}(\xi,\lambda)\right).$$

Then,

$$\frac{\partial p_i^{(1)}}{\partial \lambda}(0,\lambda) = z_{ic}^{(1)}, \tag{3.7}$$

from Lemma 3.1, as $z_{i\xi}^{(1)}(0) = 0$. Thus,

$$\begin{aligned}\frac{\partial \hat{\Pi}_{i,j}^{(1)}}{\partial \lambda}(0,0;0) &= \frac{\partial}{\partial \lambda}\left\{\frac{p_i^{(1)}(0,\lambda)}{p_i^{(2)}(0,\lambda)}\right\}\Bigg|_{\lambda=0}\\ &= \frac{p_{i\lambda}^{(1)}}{p_i^{(2)}}\Bigg|_{\lambda=0} - \frac{p_i^{(1)}p_{i\lambda}^{(2)}}{\left\{p_i^{(2)}\right\}^2}\Bigg|_{\lambda=0}\\ &= \frac{z_{ic}^{(1)}(0)}{p_i^{(2)}(0,0)} \qquad (\text{as } p_i^{(1)}(0,0) = 0)\\ &= -\frac{z_{ic}^{(1)}(0)}{\lambda_{j,s}} \qquad (\text{as } p_i^{(2)}(0,0) = z_{i\xi}^{(2)}(0) = -\lambda_{i,s}).\end{aligned} \tag{3.8}$$

The last expression is not zero from the transversality condition (2.8) and the fact that the bifurcation curve is expressed as the graph of a smooth function $c_{1,1}(\gamma)$. Moreover, the bifurcation diagram (Figure 1.2) shows that it is positive. □

4. The eigenvalue equation. With $\hat{\Pi}_i$ and $\hat{\Pi}_{i,j}$, the condition for λ being an eigenvalue is expressed as follows.

LEMMA 4.1.

$$E(\lambda) := \hat{\Pi}_{1,2}^{(1)} \circ \hat{\Pi}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\lambda) = 0 \tag{4.1}$$

if and only if λ is an eigenvalue of the linearization operator L along the travelling 1-front wave solution corresponding to the 1-heteroclinic solution from a_1 to a_2.

Proof. $E(\lambda)$ corresponds to the $p^{(1)}$ component of the intersection $\tilde{W}^u(a_1,\lambda,\gamma,c) \cap \hat{\Sigma}_{2,s}$, and this is zero if and only if $\tilde{W}^u(a_1,\lambda,\gamma,c) \subset \tilde{W}^s(a_2,\lambda,\gamma,c)$. Thus $E(\lambda) = 0$ if and only if (1.16) has a bounded solution for the p component along the 1-heteroclinic solution for the z component. This means the lemma holds. □

Let us calculate the leading terms of $E(\lambda)$. For later use, let $\tilde{\lambda} = \frac{\lambda}{\delta_2}$ and expand $E(\lambda)$ in $\tilde{\lambda}$.

First, notice that $\hat{\Pi}_{1,2}^{(1)}(0,0;0) = -\frac{\delta_2}{\Lambda_{2,s}}$, as $\Pi_{1,2}^{(1)}(0,0) = \delta_2$ and $\hat{\Pi}_{1,2}(0,0;0)$ corresponds to the tangent vector for the 1-heteroclinic solution $z_1(\xi)$ at $z_1(\xi) \in \Sigma_{2,s}$. Similarly, $\hat{\Pi}_{1,2}^{(3)}(0,0;0) = f\delta_2 + O(\delta^2)$, where $f = \frac{\partial \hat{\Pi}_{1,2}^{(3)}}{\partial \delta_2}(0,0;0)$. Thus

$$\begin{pmatrix} \hat{\Pi}_{1,2}^{(1)}(0,0;\delta_2\tilde{\lambda}) \\ \hat{\Pi}_{1,2}^{(3)}(0,0;\delta_2\tilde{\lambda}) \end{pmatrix} = \begin{pmatrix} c_2^{(1)}\delta_2\tilde{\lambda} - \frac{\delta_2}{\Lambda_{2,s}} + O(\delta_2^2\tilde{\lambda}^2) \\ f\delta_2 + c_2^{(3)}\delta_2\tilde{\lambda} + O(\delta_2^2, \delta_2^2\tilde{\lambda}^2) \end{pmatrix} \tag{4.2}$$

where $c_j^{(k)} = \frac{\partial \hat{\Pi}_{i,j}^{(k)}}{\partial \lambda}(0,0;0)$. Then,

$$\begin{pmatrix} \hat{\Pi}_2^{(2)} \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) \\ \hat{\Pi}_2^{(3)} \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) \end{pmatrix} = \begin{pmatrix} \frac{\delta_2^{\hat{\Lambda}_{2,s}}}{c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)} \\ \frac{c_2^{(3)}\tilde{\lambda} + f + O(\delta_2, \delta_2\tilde{\lambda}^2)}{c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)}\delta_2^{1+\hat{\Lambda}_{2,ss}} \end{pmatrix}. \tag{4.3}$$

Similarly,

$$\hat{\Pi}_{2,1}^{(1)} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;0) = -\frac{\delta_1^{(1)}}{\Lambda_{1,s}}, \tag{4.4}$$

$$\begin{aligned} &\hat{\Pi}_{2,1}^{(1)} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) \\ &= -\frac{\delta_1^{(1)}}{\Lambda_{1,s}} + c_1^{(1)}\tilde{\lambda} + \hat{d}_1^{(1)}\left\{\hat{\Pi}_2^{(2)} \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) - \hat{\Pi}_2^{(2)} \circ \hat{\Pi}_{1,2}(0,0;0)\right\} \\ &\qquad + \hat{e}_1^{(1)}\left\{\hat{\Pi}_2^{(3)} \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) - \hat{\Pi}_2^{(3)} \circ \hat{\Pi}_{1,2}(0,0;0)\right\} \\ &\qquad + O(\delta_2^2\tilde{\lambda}^2, \delta^{2\hat{\Lambda}_{2,s}}, \delta_2^{2+2\hat{\Lambda}_{2,ss}}, \delta_2^{1+\hat{\Lambda}_{2,s}}\tilde{\lambda}, \delta_2^{2+\Lambda_{2,ss}}\tilde{\lambda}) \\ &= -\frac{1}{\Lambda_{1,s}}\left(\frac{\delta_2}{-d_2^{(1)}}\right)^{\frac{1}{\Lambda_{1,s}}} + c_1^{(1)}\delta_2\tilde{\lambda} + \frac{\hat{d}_1^{(1)}\delta_2^{\hat{\Lambda}_{2,s}}}{c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)} + \hat{d}_1^{(1)}\Lambda_{2,s}\delta_2^{\Lambda_{2,s}} \\ &\qquad + O\left(\delta_2^{\min\{2\hat{\Lambda}_{2,s},1\}}\right), \end{aligned} \tag{4.5}$$

where $\hat{d}_j^{(k)} = \frac{\partial \hat{\Pi}_{i,j}^{(k)}}{\partial \pi_2}(0,0;0)$ and $\hat{e}_j^{(k)} = \frac{\partial \hat{\Pi}_{i,j}^{(k)}}{\partial \pi_3}(0,0;0)$, and

$$\hat{\Pi}_{2,1}^{(3)} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;0) = o(1) \qquad (\delta_2 \to 0), \tag{4.6}$$

$$\hat{\Pi}^{(3)}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) = o(1) \qquad (\delta_2 \to 0). \tag{4.7}$$

This means the following.

$$\begin{aligned}
&\hat{\Pi}^{(1)}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) \\
&= \frac{\left\{\delta_1^{(1)}\right\}^{1+\hat{\Lambda}_{1,s}}}{-\frac{1}{\Lambda_{1,s}}\left(\frac{\delta_2}{-d_2^{(1)}}\right)^{\frac{1}{\Lambda_{1,s}}} + c_1^{(1)}\delta_2\tilde{\lambda} + \frac{\hat{d}_1^{(1)}\delta_2^{\hat{\Lambda}_{2,s}}}{c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)} + \hat{d}_1^{(1)}\Lambda_{2,s}\delta_2^{\Lambda_{2,s}} + O\left(\delta_2^{\min\{2\hat{\Lambda}_{2,s},1\}}\right)} \\
&= \frac{\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)\right)\left\{\left(\frac{\delta_2}{-d_2^{(1)}}\right)^{\frac{1}{\Lambda_{1,s}}} + O\left(\delta_2^{\frac{1}{\Lambda_{1,s}}+\Delta}\right)\right\}^{1+\hat{\Lambda}_{1,s}}}{\left\{c_1^{(1)}\delta_2\tilde{\lambda} + \hat{d}_1^{(1)}\Lambda_{2,s}\delta_2^{\Lambda_{2,s}} + O\left(\delta_2^{\min\{2\hat{\Lambda}_{2,s},1\}}\right)\right\}\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + O(\delta_2\tilde{\lambda}^2)\right) + \hat{d}_1^{(1)}\delta_2^{\hat{\Lambda}_{2,s}}} \\
&= \frac{\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right)\left(\frac{1}{-d_2^{(1)}}\right)^{\frac{1+\hat{\Lambda}_{1,s}}{\Lambda_{1,s}}}\delta_2^{\frac{1+\hat{\Lambda}_{1,s}}{\Lambda_{1,s}}-\Lambda_{2,s}}(1+o(1))}{\left\{\hat{d}_1^{(1)}\Lambda_{2,s}+o(1)\right\}\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right) + \hat{d}_1^{(1)}} \qquad \text{as } \delta_2 \to 0.
\end{aligned} \tag{4.8}$$

Finally,

$$\begin{aligned}
E(\delta_2\tilde{\lambda}) = {} & c_2^{(1)}\delta_2\tilde{\lambda} \\
& + \hat{d}_2^{(1)}\left(\hat{\Pi}^{(1)}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) - \hat{\Pi}^{(1)}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;0)\right) \\
& + \hat{e}_2^{(1)}\left(\hat{\Pi}^{(3)}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) - \hat{\Pi}^{(3)}_1 \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;0)\right) \\
& + O\left(\delta_2^2\right) \\
= {} & c_2^{(1)}\delta_2\tilde{\lambda} \\
& + \hat{d}_2^{(1)}\frac{\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right)\left(\frac{1}{-d_2^{(1)}}\right)^{\frac{1+\hat{\Lambda}_{1,s}}{\Lambda_{1,s}}}\delta_2^{\frac{1+\hat{\Lambda}_{1,s}}{\Lambda_{1,s}}-\Lambda_{2,s}}(1+o(1))}{\left\{\hat{d}_1^{(1)}\Lambda_{2,s}+o(1)\right\}\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right) + \hat{d}_1^{(1)}} \\
& + \hat{d}_2^{(1)}\frac{\Lambda_{1,s}\delta_2}{-d_2^{(1)}} \\
& + \hat{e}_2^{(1)}\hat{\Pi}^{(3)}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;\delta_2\tilde{\lambda}) \\
& \times \frac{\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right)\left(\frac{1}{-d_2^{(1)}}\right)^{\frac{1+\hat{\Lambda}_{1,ss}}{\Lambda_{1,s}}}\delta_2^{\frac{1+\hat{\Lambda}_{1,ss}}{\Lambda_{1,s}}-\Lambda_{2,s}}(1+o(1))}{\left\{\hat{d}_1^{(1)}\Lambda_{2,s}+o(1)\right\}\left(c_2^{(1)}\tilde{\lambda} - \frac{1}{\Lambda_{2,s}} + o(1)\right) + \hat{d}_1^{(1)}} \\
& + \hat{e}_2^{(1)}\Lambda_{1,s}\left(\frac{\delta_2}{-d_2^{(1)}}\right)^{1+\frac{\hat{\Lambda}_{1,ss}-\hat{\Lambda}_{1,s}}{\Lambda_{1,s}}}(1+o(1)) \\
& + O\left(\delta_2^2\right)
\end{aligned} \tag{4.9}$$

as $E(0) = 0$ and

$$\begin{aligned} &\hat{\Pi}_1^{(3)} \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;*) \\ &= \hat{\Pi}_{2,1}^{(3)} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;*) \times \hat{\Pi}_1^{(1)} \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;*) \\ &\times \left(\delta_1^{(1)}\right)^{\hat{\Lambda}_{1,ss} - \hat{\Lambda}_{1,s}} . \end{aligned} \tag{4.10}$$

5. Proof of the theorem.

In this section we prove the following theorem.

THEOREM. *Assume that the system* (1.6) *is* C^r*-diffeomorphic* ($r \geq 2$) *to linear systems in some neighborhoods of equilibria* a_i*, and* ε *is small; then the travelling* 1*-front* (*back*) *wave solution of FitzHugh–Nagumo equations* (1.1) *bifurcating from simple front and back travelling wave solutions is stable.*

We give the proof for the 1-front wave; the proof for the 1-back wave is similar.

First, let L be the linearization operator with respect to the 1-front wave solution $z_1(\xi)$; then we have the following for spectrum $\sigma(L)$ of L.

PROPOSITION 5.1. *For some negative constant* β,

$$\sharp\left(\sigma(L) \cap \{\lambda | Re\lambda > \beta\}\right) = 3, \tag{5.1}$$

counting their multiplicity.

Proof. Let $z_1^*(\xi)$ be the simple front wave and $z_2^*(\xi)$ be the simple back wave, and let L_i ($i = 1, 2$) be the linearization operator with respect to them. Then it is well known (see Evans [8] or Jones [11]) that there exists a negative constant β so that $\sigma(L) \cap \{\lambda | \mathrm{Re}\lambda > \beta\}$ and $\sigma(L_i) \cap \{\lambda | \mathrm{Re}\lambda > \beta\}$ consist of isolated eigenvalues with finite multiplicity. Moreover, this β can be chosen so that $\sigma(L_i) \cap \{\lambda | \mathrm{Re}\lambda > \beta\} = \{0\}$ holds (see Yanagida [17]). Thus, simply applying the additive formula for eigenvalues (see Alexander and Jones [2] or Nii [12]), we conclude that $\sharp\left(\sigma(L) \cap \{\lambda | \mathrm{Re}\lambda > \beta\}\right) = 3$ for the 1-front wave. □

PROPOSITION 5.2. *L has even eigenvalues with positive real part.*

Proof. First, we remark that if L has an eigenvalue which is not real, then L also has an eigenvalue which is complex conjugate to it.

Let $E\mathrm{v}(\lambda)$ be the Evans function for the 1-front wave $z_1(\xi)$; *i.e.,*

$$E\mathrm{v}(\lambda) = \det\left(p_1(\xi;\lambda) p_2(\xi;\lambda) p_3(\xi;\lambda)\right)\big|_{\xi=0},$$

where $p_1(\xi;\lambda)$ is a solution of (1.13) along $z_1(\xi)$ which is bounded as $\xi \to -\infty$, and $p_2(\xi;\lambda)$ and $p_3(\xi;\lambda)$ are solutions which are bounded as $\xi \to +\infty$. This function is analytic in λ and vanishes at the eigenvalues of L, and each order of each vanishing point is equal to the multiplicity of the eigenvalue. See Alexander, Gardner, and Jones [1] for more detail. If we choose p_1 and p_2 so that $p_1(\xi;0) = p_2(\xi;0) = z_{1\xi}(\xi)$ then, by Lemmas 3.1 and 3.2, the derivative of $E\mathrm{v}(\lambda)$ with respect to λ at $\lambda = 0$ is expressed as follows:

$$\begin{aligned} \tfrac{\partial E\mathrm{v}}{\partial \lambda}(0) &= \det\left(\left(p_{1\lambda}(\xi;0) - p_{2\lambda}(\xi;0)\right) z_{1\xi}(\xi) p_3(\xi;0)\right)\big|_{\xi=0} \\ &= \det\left(z_{1c}(\xi) z_{1\xi}(\xi) p_3(\xi;0)\right)\big|_{\xi=0}. \end{aligned} \tag{5.2}$$

The same expression holds for the Evans function $E\mathrm{v}^*(\lambda)$ corresponding to the simple front $z_1^*(\xi)$. Moreover, $z_1(\xi)$ and $z_1^*(\xi)$ can be taken so that $z_{1c}(0)$ and $z_{1c}^*(0)$, $z_{1\xi}(0)$ and $z_{1\xi}^*(0)$, and $p_3(0;\lambda)$ for both cases coincide in the limit of $(\gamma, c) \to (\gamma_0, c_0)$. Thus the signs of $\frac{\partial E\mathrm{v}}{\partial \lambda}(0)$ and $\frac{\partial E\mathrm{v}^*}{\partial \lambda}(0)$ agree.

Here if $E\mathrm{v}$ and $E\mathrm{v}^*$ are normalized so that they are positive for large λ, then $\frac{\partial E\mathrm{v}^*}{\partial\lambda}(0) > 0$ because the simple front is stable (Yanagida [17]) and $E\mathrm{v}^*(\lambda)$ does not vanish for $\lambda > 0$. This means $\frac{\partial E\mathrm{v}}{\partial\lambda}(0) > 0$, so $E\mathrm{v}(\lambda) = 0$ for even $\lambda > 0$; thus the proposition holds. □

By the propositions above, for proof of the stability, it suffices to find one eigenvalue with negative real part which is near 0. This is achieved by proving existence of a negative solution of (4.1).

First, the right-hand side of (4.3) tends to zero as δ_2 tends to zero, provided that $\tilde{\lambda}$ is negative. Similarly, the last expression of (4.8) converges to zero uniformly in $\tilde{\lambda} < -\tilde{\lambda}_0$ for arbitrary small $\tilde{\lambda}_0 > 0$, and at the same time

$$\hat{\Pi}_1^{(1)} \circ \hat{\Pi}_{2,1} \circ \hat{\Pi}_2 \circ \hat{\Pi}_{1,2}(0,0;0) \to 0 \qquad (\delta_2 \to 0). \tag{5.3}$$

Therefore, the expansion (4.5) and (4.8) is valid for $\tilde{\lambda} < -\tilde{\lambda}_0$ when δ_2 tends to zero.

Here, (4.9) implies

$$E(\delta_2\tilde{\lambda}) = c_2^{(1)}\delta_2\tilde{\lambda} + \hat{d}_2^{(1)}\frac{\Lambda_{1,s}\delta_2}{-d_2^{(1)}} + o(\delta_2) \qquad (\delta_2 \to 0). \tag{5.4}$$

Thus

$$\frac{1}{\delta_2}E(\delta_2\tilde{\lambda}) \to c_2^{(1)}\tilde{\lambda} - \frac{\hat{d}_2^{(1)}}{d_2^{(1)}}\Lambda_{1,s} \quad \text{as} \quad \delta_2 \to 0. \tag{5.5}$$

This means $\frac{1}{\delta_2}E(\delta_2\tilde{\lambda}) = 0$ has a solution $\tilde{\lambda} = \hat{d}_2^{(1)}\Lambda_{1,s}/d_2^{(1)}c_2^{(1)} + O(\delta_2)$ for small $\delta_2 > 0$; i.e., $E(\lambda) = 0$ has a solution $\lambda = (\hat{d}_2^{(1)}\Lambda_{1,s}/d_2^{(1)}c_2^{(1)})\delta_2 + o(\delta_2)$. By the inequality (2.7) and Lemma 3.3, $c_2^{(1)} > 0$, $d_2^{(1)} < 0$, and $\hat{d}_2^{(1)} > 0$, so this solution is negative. As a result, the 1-front wave solution is stable by the argument above. This completes the proof of the theorem.

Appendix. Proof of Proposition 2.2. Let $\mathcal{V}(n)$ be the vector space of germs of the C^∞-vector fields at 0 in $\mathbb{R}^n$ and let $\mathcal{E}(n)$ be the algebra of C^∞-germs of functions on $\mathbb{R}^n$ at 0. The maximal ideal of $\mathcal{E}(n)$ generated by arbitrary C^∞-functions $f_1, \dots, f_k$ shall be denoted by $\mathcal{M}(f_1, \dots, f_k)$, and a subspace $\mathcal{V}(n; s)$ of $\mathcal{V}(n+s)$ shall be defined by

$$\mathcal{V}(n;s) = \mathcal{M}(x_1, \dots, x_n)\mathrm{span}\left\{\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\right\}. \tag{A.1}$$

For the sake of convenience we write the coordinate of $\mathbb{R}^{n+s}$ as $x_1, \dots, x_n, y_1, \dots, y_s$. The vector field $\mathrm{pr}_1 \circ X(x, y)$ is expressed as X_y for a given vector field $X \in \mathcal{V}(n; s)$, where pr_1 is the projection onto the first n coordinates. Hence $X_y \in \mathcal{V}(n)$ for every $y \in \mathbb{R}^s$ sufficiently small.

Let $\lambda_1(y), \dots, \lambda_n(y)$ be the eigenvalues of $DX_y(0)$ and assume that $\mathrm{Re}\lambda_i > 0$ for $i = 1, \dots, p$ and $\mathrm{Re}\lambda_i < 0$ for $i = p+1, \dots, n$.

We introduce two constants depending on eigenvalues:

$$\begin{aligned} A_+ &= \frac{\max\limits_{p+1\le i\le n} |\mathrm{Re}\lambda_i(0)|}{\min\limits_{1\le i\le p} |\mathrm{Re}\lambda_i(0)|}, \\ A_- &= \frac{\max\limits_{1\le i\le p} |\mathrm{Re}\lambda_i(0)|}{\min\limits_{p+1\le i\le n} |\mathrm{Re}\lambda_i(0)|}, \end{aligned} \tag{A.2}$$

and let $k_\pm$ be two integers which satisfy the condition below for some integer r:

$$k_\pm > r(1 + A_\pm). \tag{A.3}$$

Then, the following holds.

PROPOSITION A.1 (Rychlic [14]). *Let $X \in \mathcal{V}(n; s)$ and $k = k_+ + k_-$, and let X' be the Taylor expansion of X of degree $k - 1$. Then X and X' are C^r-equivalent.*

By this proposition X can be linearized through a C^r-coordinate change if the Taylor expansion of X vanishes up to degree $k - 1$.

Here, by the Poincaré–Dulac theorem (see, for example, Arnol'd [4]), this can be achieved through polynomial transformation provided that $\{\lambda_1(y), \dots, \lambda_n(y)\}$ satisfies nonresonant condition up to order $k-1$. Thus the proof of the Proposition 2.2 amounts to the proof the following lemma.

LEMMA A.1. *There are uncountably many pairs of (a, ε) $(0 < a < \frac{1}{2}, 0 < \varepsilon < \varepsilon_0)$ such that for all $k > 0$, $\{\lambda_{i,u}(\mu), -\lambda_{i,s}(\mu), -\lambda_{i,ss}(\mu)\}$ satisfies a nonresonant condition up to order k for $\mu = (\gamma(\varepsilon), c(\varepsilon), \varepsilon)$ and $(i = 1, 2)$.*

Proof. The bifurcation point $(\gamma(\varepsilon), c(\varepsilon))$ in the γ–c plane is smooth in ε and thus $\lambda_{i,u}(\gamma(\varepsilon), c(\varepsilon), \varepsilon), -\lambda_{i,s}(\gamma(\varepsilon), c(\varepsilon), \varepsilon), -\lambda_{i,ss}(\gamma(\varepsilon), c(\varepsilon), \varepsilon)$ $(i = 1, 2)$ are also smooth in ε. Moreover, $\lambda_{i,u}(\gamma(0), c(0), 0) = \frac{1}{\sqrt{2}}$, $-\lambda_{i,s}(\gamma(0), c(0), 0) = 0$, $-\lambda_{i,ss}(\gamma(0), c(0), 0) = -\sqrt{2}a$ $(i = 1, 2)$, and $-\lambda_{i,s}(\gamma(\varepsilon), c(\varepsilon), \varepsilon) < 0$ for small $\varepsilon > 0$.

Notice that the system (1.6) is symmetric with respect to the coordinate change

$$(u, v, w) \mapsto \left(-u + \frac{2(a+1)}{3}, -v, -w + \frac{2(2-a)(1-2a)(a+1)}{27}\right) \tag{A.4}$$

for $\gamma = \frac{9}{(2-a)(1-2a)} = \gamma(0)$ for any $\varepsilon > 0$. Then $\gamma(\varepsilon) = \gamma(0)$ for $\varepsilon > 0$ as $(\gamma(\varepsilon), c(\varepsilon))$ is the only parameter value at which $z_1^*(\xi)$ and $z_2^*(\xi)$ coexist. Thus $\lambda_{1,u} = \lambda_{2,u}$, $-\lambda_{1,s} = -\lambda_{2,s}$, and $-\lambda_{1,ss} = -\lambda_{2,ss}$ holds for $(\gamma(\varepsilon), c(\varepsilon))$ because of the symmetry.

For each $m = (m_u, m_s, m_{ss}) \in \mathbb{Z}^3$ $(m_* \geq 0,\ m_u + m_s + m_{ss} \geq 2)$, let I_m^u be a union of intervals such that $\varepsilon \in I_m^u$ if and only if

$$\lambda_{i,u} = m_u \lambda_{i,u} + m_s(-\lambda_{i,s}) + m_{ss}(-\lambda_{i,ss}). \tag{A.5}$$

Similarly, $\varepsilon \in I_m^s$ if and only if

$$-\lambda_{i,s} = m_u \lambda_{i,u} + m_s(-\lambda_{i,s}) + m_{ss}(-\lambda_{i,ss}) \tag{A.6}$$

and $\varepsilon \in I_m^{ss}$ if and only if

$$-\lambda_{i,ss} = m_u \lambda_{i,u} + m_s(-\lambda_{i,s}) + m_{ss}(-\lambda_{i,ss}) \tag{A.7}$$

for $\varepsilon \geq 0$. Then $I_m^* = \coprod_j I_m^{*(j)}$ $(* = u, s, ss)$, where $I_m^{*(j)}$ is a closed interval, and if $\mathrm{int}(I_m^{*(j)}) \neq \emptyset$ and $\mathrm{int}(I_{m'}^{*(j')}) \neq \emptyset$ then $I_m^{*(j)} \cap I_{m'}^{*(j')} = \emptyset$ because $\lambda_{i,u}$, $-\lambda_{i,s}$, and $-\lambda_{i,ss}$ are smooth in ε.

At $\varepsilon = 0$, the only resonances are of the form

$$-\lambda_{i,s}(\gamma(0), c(0), 0) = m\{-\lambda_{i,s}(\gamma(0), c(0), 0)\} \qquad (m \geq 2); \tag{A.8}$$

i.e., $-\lambda_{i,s}(\gamma(0), c(0), 0) = 0$ if a is irrational. This means there is no interval with $0 \in I_m^{*(j)}$ and $\mathrm{int}(I_m^{*(j)}) \neq \emptyset$. Thus, for each fixed irrational a, there exist infinitely many ε such that none of (A.5), (A.6), or (A.7) holds for any $m = (m_u, m_s, m_{ss})$. This completes the proof. □

REFERENCES

[1] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J. Alexander and C. Jones, *Existence and stability of asymptotically oscillatory double pulses,* J. Reine Angew. Math., 446 (1994), pp. 49–79.

[3] J. Alexander and C. Jones, *Existence and stability of asymptotically oscillatory triple pulses,* Z. Angew. Math. Phys., 44 (1993), pp. 189–200.

[4] V. Arnol'd, *Geometrical methods in the theory of ordinary differential equation,* English trans., M. Levi, ed., Springer-Verlag, New York, 1983.

[5] P. Bates and C. Jones, *Invariant manifolds for semilinear partial differential equations,* Dynamics Reported, 2 (1988), pp. 1–38.

[6] B. Deng, *The bifurcations of countable connections from a twisted heteroclinic loop,* SIAM J. Math. Anal., 22 (1991), pp. 653–678.

[7] B. Deng, *The existence of infinitely many travelling front and back waves in the FitzHugh–Nagumo equations,* SIAM J. Math. Anal., 22 (1991), pp. 1631–1650.

[8] J. Evans, *Nerve axon equations,* III: *Stability of the nerve impulse,* Indiana Univ. Math. J., 22 (1972), pp. 577–594.

[9] J. Evans, *Nerve axon equations,* IV: *The stable and unstable impulses,* Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[10] D. Henry, *The geometric theory of semilinear parabolic equations,* Lec. Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[11] C. Jones, *Stability of the travelling wave solution of the FitzHugh–Nagumo system,* Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[12] S. Nii, *An extension of the stability index for travelling wave solutions and its application for bifurcations,* SIAM J. Math. Anal., 28 (1997), pp. 402–433.

[13] S. Nii, *A topological proof of stability of N-front solutions of the FitzHugh–Nagumo equations,* preprint.

[14] M. Rychlic, *Lorenz attractors through Šil'nikov-type bifurcation. Part* I, Ergod. Th. & Dynam. Sys., 10 (1989), pp. 793–821.

[15] B. Sandstede, *Stability of N-fronts bifurcating from a twisted heteroclinic loop and an application to the FitzHugh–Nagumo equation,* SIAM J. Math. Anal., to appear.

[16] E. Yanagida, *Stability of fast travelling pulse solutions of the FitzHugh–Nagumo equations,* J. Math. Biology, 22 (1985), pp. 81–104.

[17] E. Yanagida, *Stability of travelling front solutions of the FitzHugh–Nagumo equations,* Math. Comput. Modelling, 12 (1989), pp. 289–301.

[18] E. Yanagida and K. Maginu, *Stability of double-pulse solutions in nerve axon equations,* SIAM J. Appl. Math., 49 (1989), pp. 1158–1173.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1113–1134, September 1997

007

ON THE ASYMPTOTIC SOLUTION OF LAMINAR CHANNEL FLOW WITH LARGE SUCTION*

CHUNQING LU†

Abstract. The equation considered is

$$\epsilon f^{iv} = ff''' - f'f'',$$

with boundary conditions

$$f(0) = f''(0) = 0, f(1) = 1, f'(1) = 0.$$

When $0 < \epsilon \ll 1$, the boundary value problem corresponds to the laminar flow of a viscous fluid through a porous channel under large suction. It is known that there are three solutions in this case: two of them are monotone increasing (types I and II), and the third is nonmonotone (type III). Let $(1-\Delta)$ be the turning point of $f(\eta)$ in (0, 1). This paper presents a rigorous proof of the asymptotic behavior of type III solutions, which is

$$f(\eta) \sim \kappa \sin \frac{\pi\eta}{1-\Delta}, \quad \text{where} \quad \kappa \sim \frac{1-\Delta}{\pi\Delta} \text{ and } \frac{\Delta}{\epsilon} e^{\Delta/\epsilon} \sim \frac{1}{2e\pi^9\epsilon^8},$$

uniformly on $[0, 1-\Delta]$ as $\epsilon \to 0^+$, and provides detailed information at the turning point.

Key words. laminar flow, turning point, exponential terms

AMS subject classification. 34B15

PII. S0036141096297704

1. Introduction.

Assuming that the fluid is incompressible and is injected or sucked through the walls of a rectangular channel, Berman [1], in 1953, reduced the boundary value problem for the Navier–Stokes equations to a nonlinear fourth-order ordinary differential equation

$$\epsilon f^{iv} = ff''' - f'f''. \tag{1.1}$$

If the flow is assumed to be symmetric, then f is an odd function and the boundary conditions

$$f(0) = f''(0) = 0, \tag{1.2}$$
$$f'(1) = 0, f(1) = 1 \tag{1.3}$$

are imposed, where $f(\eta)$ is the unknown function related to the stream function and η is the normalized transverse coordinate ($\eta = \pm 1$ are the walls). The parameter ϵ equals $\frac{1}{R}$ where R is the Reynolds number of the flow. The case $\epsilon < 0$ corresponds to the injection problem arising in transpiration cooling, and $\epsilon > 0$ to the suction problem for the isotope separation.

The boundary value problem (1.1)–(1.2)–(1.3) admits at least three solutions, which were classified as type I, the increasing concave down function, type II, the increasing function with a reflection point, and type III, the nonmonotone function with a turning point [12]. Here the turning point, denoted by $z_\epsilon = 1-\Delta_\epsilon$, is defined as

*Received by the editors January 26, 1996; accepted for publication May 29, 1996.
http://www.siam.org/journals/sima/28-5/29770.html

†Department of Mathematics and Statistics, Southern Illinois University at Edwardsville, Edwardsville, IL 62026 (clu@daisy.ac.siue.edu).

the value of $\eta \in (0,1)$ at which the solution vanishes. There have been many published numerical results and formally asymptotic analyses [2], [3], [8], [10], [11], [13], [14], [15]; see [7] and the references listed there. The existence of the three solutions for sufficiently small $\epsilon > 0$ can be found in [4]. As far as rigorous asymptotic results, Hastings, Lu, and MacGillivray studied the behavior of type I and type II solutions as $|R| \to \infty$ in [4] and [6], and McLeod proved the asymptotic formula of $f'''(0)$ for type I and type II solutions [9]. As for type III solutions, Robinson [12], Zaturska, Drazin, and Banks [16], and MacGillivray and Lu [7] presented formal asymptotic results in different ways. But from the rigorous analysis point of view, only very little was known about type III solutions. This paper presents a rigorous proof of the asymptotic formula for type III solutions of (1.1)–(1.2)–(1.3) as $\epsilon \to 0^+$.

The paper is organized as follows. Section 2 contains the main result of the paper. In section 3, we study the function $g(\eta, \epsilon) \equiv \frac{f(\eta,\epsilon)}{f'(0,\epsilon)}$ for $\eta \in [0, z_\epsilon]$, which turns out to behave like a sine function on that interval. This implies that the investigation of $f'(0, \epsilon)$ is crucial. In section 4, another rescaled function $u(t, \epsilon^*) = f(z_\epsilon + \Delta_\epsilon t)$, where $t = \frac{\eta - z_\epsilon}{\Delta_\epsilon}$ and $\epsilon^* = \frac{\epsilon}{\Delta_\epsilon}$, is introduced. Using the boundary conditions at $t = 1$ ($\eta = 1$) and $t = 0$ ($\eta = z_\epsilon$), we prove that as $\epsilon \to 0^+$, (1) $\epsilon^* \to 0^+$, (2) $|f'(0, \epsilon)|\Delta_\epsilon \to 1$, and (3) the limit function of $u(t, \epsilon^*)$ for $t \in [0, 1]$ is linear. Since $u'''(0) \sim \Delta_\epsilon^2 \pi^2$, it is necessary to approximate $u'''(0)$ in order to find the asymptotic value for $f'(0, \epsilon)$. In section 5, we prove the asymptotic behavior of $u(t)$ for $t > 0$, including the boundary layer on the right side, as well as one asymptotic formula linking $u'''(0)$ and $u^{iv}(0)$. In section 6, we return to the left side of the turning point and prove, with the results from section 1, another asymptotic formula linking $u'''(0)$ and $u^{iv}(0)$. In section 7, the leading terms for $u'''(0)$, $u^{iv}(0)$, and $u''(0)$ are determined. These and the asymptotic formula of $u'(0)$ given in section 5 completely provide the asymptotic behavior of the solution at the turning point. The asymptotic relation between Δ_ϵ and ϵ is finally determined at the end of section 7.

Since some previous rigorous results on the nonmonotone solutions are applied repeatedly, they should be listed here.

(I) $f^{iv}(\eta) < 0$ for all $\eta \in (0, 1]$.

(II) For any given sufficiently small $\epsilon > 0$, a type III solution exists, which means that there exists a pair $\alpha < 0$ and $\beta > 0$ depending on ϵ such that the initial value problem of (1.1) with (1.2) and $f'(0) = \alpha, f'''(0) = \beta$ has a solution satisfying (1.3).

(III) Let $z_\epsilon \equiv 1 - \Delta_\epsilon$ be the turning point of the solution, and y_ϵ the reflection point of $f(\eta)$. Then, $y_\epsilon > z_\epsilon$, and hence, $f'' > 0$ in $(0, y_\epsilon\ )$ and $f'' < 0$ in $(y_\epsilon, 1)$. Also, $f'(z_\epsilon) > 0, f''(z_\epsilon) > 0$, $z_\epsilon \to 1$, and $y_\epsilon \to 1$ as $\epsilon \to 0^+$.

(IV) As $\epsilon \to 0^+$, $\alpha \to -\infty$, $f(\eta) \to -\infty$, and $f''(\eta) \to \infty$ for each $\eta \in (0, 1)$.

Proofs of (I)–(IV) can be found in [4] and [7]. Since a type III solution exists for only sufficiently small $\epsilon > 0$, throughout the rest of the paper the expression $\epsilon \to 0$ means $\epsilon \to 0^+$. Also, for briefness, the dependence on ϵ is often dropped, such as $f(\eta, \epsilon) \equiv f(\eta)$, $g(\eta, \tilde{\epsilon}) \equiv g(\eta), h(\eta, \tilde{\epsilon}) \equiv h(\eta), \Delta_\epsilon \equiv \Delta$, $u(t, \epsilon^*) \equiv u(t)$, and so on, except for cases where the appearance of ϵ is necessary.

2. Main result.

The main result of the paper is stated in the following theorem.

THEOREM 2.1. *As $\epsilon \to 0^+$, the nonmonotone solution $f(\eta, \epsilon)$ of* (1.1)–(1.2)–(1.3) *satisfies*

$$f(\eta, \epsilon) \sim -\frac{1 - \Delta}{\pi\Delta} \sin \frac{\pi\eta}{1 - \Delta}$$

uniformly on $[0, 1-\Delta]$, *where* $f(1-\Delta, \epsilon) = 0$, *and* Δ *satisfies*

$$\frac{\Delta}{\epsilon} e^{\frac{\Delta}{\epsilon}} \sim \frac{1}{2e\pi^9\epsilon^8}.$$

Note from Theorem 2.1 that the domain of validity for the asymptotic solution includes the moving right endpoint of the interval. The uniform convergence on the closed moving interval, which is needed in the paper, is defined in the same way that one defines uniform convergence for a fixed closed interval [5]. The theorem, of course, implies that the asymptotic formula for $f(\eta, \epsilon)$ holds not only uniformly on any compact subinterval of $[0, 1-\Delta)$, but also at the turning point $\eta = 1-\Delta$. Let $\phi(\eta, \epsilon) \equiv \frac{1-\Delta}{\pi\Delta} \sin \frac{\pi\eta}{1-\Delta}$. The first asymptotic formula of the theorem means that $\lim_{\epsilon\to 0} \frac{f(\eta,\epsilon)}{\phi(\eta,\epsilon)} = 1$ uniformly for $\eta \in [0, 1-\Delta]$. Since $f(0, \epsilon) = h\,(0, \epsilon) = 0$ for all ϵ, the ratio $\frac{f(\eta,\epsilon)}{\phi(\eta,\epsilon)}$ evaluated at $\eta = 0$ is defined, as usual, by $\lim_{\eta\to 0^+} \frac{f(\eta,\epsilon)}{\phi(\eta,\epsilon)}$, and hence is equal to $\frac{f'(0,\epsilon)}{\phi'(0,\epsilon)}$. Similarly, at $\eta = 1-\Delta$, the ratio is defined as $\lim_{\eta\to(1-\Delta)^-} \frac{f(\eta,\epsilon)}{\phi(\eta,\epsilon)} = \frac{f'(1-\Delta,\epsilon)}{\phi'(1-\Delta,\epsilon)}$.

The theorem is a consequence of Theorems 3.1, 4.1, and 4.5 in the remainder of the paper.

3. Approximation on $[0, z_\epsilon]$. Because of the unboundedness of the solution and the singularity of the boundary value problem, it is necessary to rescale the solution $f(\eta)$. Since $f'(0) = \alpha \to -\infty$ as $\epsilon \to 0$, we first introduce

$$g(\eta) \equiv \frac{f(\eta)}{|\alpha|} \quad \text{and } \tilde{\epsilon} \equiv \frac{\epsilon}{|\alpha|}.$$

Equation (1.1) then takes the form

$$\tilde{\epsilon} g^{iv} = gg''' - g'g'', \tag{3.1}$$

with initial conditions

$$g(0) = g''(0) = 0, \quad g'(0) = -1, \quad g'''(0) = \frac{\beta}{|\alpha|}. \tag{3.2}$$

It is clear that the process $\epsilon \to 0$ implies $\tilde{\epsilon} \to 0$. In this section, the following theorem will be proved.

THEOREM 3.1. *Let* $h(\eta) = -\frac{\sin(\pi\eta/(1-\Delta))}{\pi/(1-\Delta)}$. *Then* $\frac{f(\eta)}{|\alpha|} - h \to 0$ *in* C^4 *on* $[0, 1-\Delta]$, *and* $\frac{f(\eta)}{|\alpha|} \sim h$ *uniformly on* $[0, 1-\Delta]$ *as* $\tilde{\epsilon} \to 0$.

The proof of Theorem 3.1 consists of Lemmas 3.3, 3.5, 3.7, 3.8, and 3.10–3.12.

Throughout the paper the following two propositions are applied repeatedly.

PROPOSITION 3.2. *Let* $\{y(x, s)\}$ *be a sequence of* C^1 *functions defined on* $[a, b]$ *with* $a = a(s), b = b(s)$, *and* $[a, b] \subseteq [c, d]$ *for all* $s \in (0, 1)$ *where* s *is a parameter. Assume that* $y(x, s)$ *is uniformly bounded on* $[a, b]$. *Then,* (1) *if* $y'(x, s)$ *is nonnegative and concave down on* $[a, b]$ *for all* s, *then* $y'(x, s)$ *is uniformly bounded on* $[a, b]$; (2) *if* $y'(x, s)$ *is nonpositive and concave up on* $[a, b]$ *for all* s, *then* $y'(x, s)$ *is uniformly bounded on* $[a, b]$.

Proof. We only give a proof of (1), since the proof of (2) is similar. Suppose, for contradiction, that the sequence $\{y'(x, s)\}$ is not uniformly bounded. Then there would be a subsequence of $\{s\}$, say $\{s_n\}$, and a sequence $\{x_n\} \subseteq [a, b]$ such that

$y'(x_n,s_n)=\max\{y'(x,s_n)\}\to\infty$ as $n\to\infty$. Without loss of generality, one can assume $x_n\to p\in[c,d]$ as $n\to\infty$. Then the concavity of y' leads to, for sufficiently large integer n, either $y'(x,s_n)>\frac{y'(x_n,s_n)-y'(a,s_n)}{x_n-a}(x-a)+y'(a,s_n)$ for $x\in(a,x_n)$ if $p\neq a$, or $y'(x,s_n)>\frac{y'(x_n,s_n)-y'(b,s_n)}{b-x_n}(b-x)+y'(b,s_n)$ for $x\in(x_n,b)$ if $p=a$. In the former case, integrating the first inequality over (a,x_n), we see that $y(x_n,\epsilon_n)$ becomes unbounded as $n\to\infty$. Similarly, in the latter case $y(b,s_n)$ becomes unbounded as $n\to\infty$. These contradict the uniform boundedness of $\{y(x,\epsilon)\}$. □

Lemma 3.3. *For any given positive $\delta\ll 1$, $g-(-(\sin\frac{\pi\eta}{z_\epsilon}/\frac{\pi}{z_\epsilon}))\to 0$ in C^1 on $[0,1-\delta]$ and in C^2 on $[\delta,1-\delta]$ as $\tilde{\epsilon}\to 0$.*

Proof. Noting that $g''>0$ on $(0,z_\epsilon)$, $g'(0)=-1$, and $g'>0$ on $[z_\epsilon,1)$, one sees that $g'>-1$; hence $g\geq-\eta$ for $\eta\in(0,z_\epsilon)$, and $0\leq g\leq\frac{1}{|\alpha|}$ for $\eta\in[z_\epsilon,1]$. Thus, $g(\eta)$ is uniformly bounded on $[0,1]$ as $\tilde{\epsilon}\to 0$. Let a positive number $\delta\ll 1$ be fixed. Since $g''>0$ on $[0,z_\epsilon]$ (in particular, on $[1-\delta,1-\frac{\delta}{2}]$), $g'(1-\delta)$ must be bounded as $\tilde{\epsilon}\to 0$; otherwise, $g(1-\frac{\delta}{2})$ would be unbounded. Using the boundedness of g' and the concavity of g'' ($g^{iv}<0$), we conclude that g'' is uniformly bounded on $[0,1-\delta]$ by Proposition 3.2. Applying the property $g^{iv}<0$, we obtain that g''' is uniformly bounded on $[\delta,1-\delta]$ (otherwise, g'' would not be bounded uniformly on $[\frac{\delta}{2},\delta]$).

Now, let us consider any sequence $\{\tilde{\epsilon}_n\}$ of $\tilde{\epsilon}$ with $\tilde{\epsilon}_n\to 0$. Since g,g', and g'' are all uniformly bounded on $[0,1-\delta]$, the Arzela–Ascoli theorem can be applied to conclude that there exist a subsequence, again denoted by $\{g(\eta,\tilde{\epsilon}_n)\}$, and a function $\tilde{h}(\eta)$ such that $g(\eta,\tilde{\epsilon}_n)\to\tilde{h}(\eta)$ in $C^1_{[0,1-\delta]}$ and $g''(\eta,\tilde{\epsilon}_n)\to\tilde{h}''(\eta)$ uniformly on $[\delta,1-\delta]$ as $n\to\infty$. Our first goal is to show $\tilde{h}(\eta)=-\frac{\sin(\pi\eta)}{\pi}$.

An integration of (3.1) with respect to η yields

$$\tilde{\epsilon}g'''=gg''-(g')^2+1+\tilde{\epsilon}\tilde{\beta}, \tag{3.3}$$

where $\tilde{\beta}=\frac{\beta}{|\alpha|}>0$. Applying the uniform boundedness of g''' on $[\delta,1-\delta]$, together with the Arzela–Ascoli theorem, one sees that $g''(\eta,\tilde{\epsilon}_n)\to\tilde{h}''(\eta)$ on that interval and $\tilde{h}$ satisfies

$$\tilde{h}\tilde{h}''-(\tilde{h}')^2+c=0, \tag{3.4}$$

where $c=\lim_{\tilde{\epsilon}\to 0}(1+\tilde{\epsilon}\tilde{\beta})\geq 1$ is a constant. Since δ is arbitrarily given, we see that $\tilde{h}(\eta)$ is defined in $(0,1)$. From (3.4),

$$\tilde{h}''=\frac{(\tilde{h}')^2-c}{\tilde{h}}.$$

Then $\tilde{h}'''$ is well defined for any $\eta\in(0,1)$. A differentiation of (3.4) shows

$$\tilde{h}\tilde{h}'''-\tilde{h}'\tilde{h}''=0, \tag{3.5}$$

for $\eta\in(0,1)$. Solving (3.5) for $\eta\in[\delta,1-\delta]$, one obtains $\tilde{h}''=A\tilde{h}$ which is valid for any $\eta\in[\delta,1-\delta]$, where A is a constant independent of δ. Thus, $\tilde{h}''$ is well defined at $\eta=0$ and $\tilde{h}''(0)=0$. It then follows from (3.4) that $c=1$. All possible solutions of equation $\tilde{h}''=A\tilde{h}$ with $\tilde{h}(0)=0,\tilde{h}'(0)=-1$ are

$$\tilde{h}(\eta)=-\tilde{h},\quad \tilde{h}(\eta)=-\frac{\sin(a\eta)}{a},\quad \tilde{h}(\eta)=-\frac{\sinh(a\eta)}{a},$$

where $a=\sqrt{|A|}>0$ is a constant. Considering the fact that $\tilde{h}''$ must have the same concavity as g'' does, we may eliminate the third possibility. Suppose $\tilde{h}=-\eta$.

Then, $g' \approx -1, g'' \approx 0$ in $[\delta, 1-\delta]$. Applying the mean value theorem on $(\delta, 1-\delta)$, we see that $g''' \approx 0$ for some $\eta_0 < 1-\delta$. Since $g'(1) = 0$ and $g'(1-\delta) \approx -1$, there exists an $\eta_1 \in (1-\delta, 1)$ such that $g''(\eta_1) \approx \frac{1}{\delta}$ by the mean value theorem. Hence, $g'''(\eta_2) = \frac{g''(\eta_1)-g''(1-\delta)}{\eta_1-(1-\delta)} > \frac{1}{2\delta^2}$ for some $\eta_2 \in (1-\delta, \eta_1)$. This implies that $g^{iv} > 0$ for some $\eta < 1$, a contradiction. Therefore, $\tilde{h}(\eta) = -\frac{\sin(a\eta)}{a}$. Next, we claim $a = \pi$. Recall that $\tilde{h}$ is the limit function of a subsequence of $\{g(\eta, \tilde{\epsilon})\}$ on $[0, 1-\delta]$ and that the value of a does not depend upon the choice of δ. It is then observed that $a \leq \pi$; otherwise, $g(1-\Delta) < 0$ for sufficiently small $\tilde{\epsilon}$, a contradiction. Similarly, it is impossible to have $a < \pi$. Thus, $a = \pi$.

Summarizing the above discussion, we have shown that any sequence $\{g(\eta, \tilde{\epsilon}_n)\}$ has a subsequence converging to $\tilde{h}$ in C^1 on $[0, 1-\delta]$ and in C^2 on $[\delta, 1-\delta]$ for any given positive $\delta \ll 1$. Since $g(1-\Delta) = 0$ and $\frac{\sin(\pi\eta)}{\pi} / \frac{\sin[\pi\eta/(1-\Delta)]}{\pi/(1-\Delta)} \to 1$ uniformly on $[0, 1-\Delta]$ as $\Delta \to 0$, we can set the limit function $h(\eta) = h(\eta, \tilde{\epsilon}) = -\frac{\sin[\pi\eta/(1-\Delta)]}{\pi/(1-\Delta)}$ for the purpose of asymptotic analysis. Since $\{g(\eta, \tilde{\epsilon}_n)\}$ is arbitrarily chosen, we conclude that $g - h \to 0$ in $C^1_{[0,1-\delta]}$ and $C^2_{[\delta,1-\delta]}$ as $\tilde{\epsilon} \to 0$. The proof of Lemma 3.3 is complete. □

PROPOSITION 3.4. *Let $g(\eta, \tilde{\epsilon})$ be a solution of* (3.1) *and $k \geq 2$ be an integer. If $g, g', g'', ..., g^{(k+1)}$ are all uniformly bounded on any compact subinterval $[0, 1-\delta]$ of $[0, 1)$, then there exists a unique C^k function $\tilde{h}(\eta)$ defined on $[0, 1)$ such that $g \to \tilde{h}$ in $C^k_{[0,1-\delta]}$ as $\tilde{\epsilon} \to 0$.*

Proof. If $k = 2$, the proof of the proposition is similar to that of Lemma 3.3. For $k > 2$, a simple application of mathematical induction derives the proposition. □

At this moment, Lemma 3.3 does not automatically imply $g \sim h(\eta)$ (hence $f(\eta, \epsilon) \sim \alpha \frac{\sin[\pi\eta/(1-\Delta)]}{\pi/(1-\Delta)}$) uniformly on any compact subinterval of $[0, 1-\Delta)$. This is because, from the proof of Lemma 3.3, the choice of the convergent subsequences may depend on δ. Therefore, to obtain the desired asymptotic result, more work is needed.

An interesting question arising from Proposition 3.4 is this: what is the largest number that k may assume? In other words, how close are g and h? An inspection of g and h at $\eta = 0$ shows that $g^v(0) = 0$ and $h^v(0) \approx -\pi^4$. Therefore, $k \leq 4$. In fact, we can prove the maximum value of k is 4. To do so, we need the uniform boundedness of higher derivatives of $g(\eta)$ on [0,1 − Δ].

LEMMA 3.5. *g''' is uniformly bounded and $g - h \to 0$ in C^2 on $[0, 1-\delta]$ for any given positive $\delta \ll 1$.*

Proof. We first prove that $g^v < 0$ as long as $g' < 0$ and $g''' > 0$; i.e., g''' is positive and concave down as long as $g' < 0$, so Proposition 3.2 can be applied. Since $g' \to -\cos \pi\eta$ on $[0, 1-\delta]$ uniformly as $\tilde{\epsilon} \to 0$, g' and h' must have the same concavity for sufficiently small $\tilde{\epsilon} > 0$. Noting that h''' has exactly one zero, $\eta = \frac{1-\Delta}{2}$, which is the reflection point of the function h', we see that g' must have a unique reflection point $\eta_3 \to \frac{1}{2}$ as $\tilde{\epsilon} \to 0$. Also, g' is concave up for $\eta < \eta_3$ and concave down for $\eta > \eta_3$. The uniqueness of η_3 follows from $g^{iv} < 0$. Differentiating (3.1) with respect to η once yields

$$\tilde{\epsilon} g^v = g g^{iv} - (g'')^2, \tag{3.6}$$

which can be rewritten as

$$\left(e^{-\frac{1}{\tilde{\epsilon}} \int g(s)ds} g^{iv}\right)' = -\frac{1}{\tilde{\epsilon}} (g'')^2 e^{-\frac{1}{\tilde{\epsilon}} \int g(s)ds}. \tag{3.7}$$

Integrating (3.7) and multiplying the exponential function on both sides leads to

$$g^{iv}(\eta) = -\frac{1}{\tilde{\epsilon}}\int_0^\eta (g''(s))^2 e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta g(r)dr}ds. \tag{3.8}$$

Differentiating (3.1) with respect to η twice gives

$$\tilde{\epsilon}g^{vi} = gg^v + g'g^{iv} - 2g''g''', \tag{3.9}$$

from which

$$g^v(\eta) = \frac{1}{\tilde{\epsilon}}\int_0^\eta (g'g^{iv} - 2g''g''')e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta g(r)dr}ds. \tag{3.10}$$

The same technique applied to (3.1) shows

$$g'''(\eta) = -\frac{1}{\tilde{\epsilon}}\int_0^\eta g'g''e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta g(r)dr}ds + g'''(0)e^{\frac{1}{\tilde{\epsilon}}\int_0^\eta g(s)ds}. \tag{3.11}$$

In order to determine the sign of g^{iv} with (3.10), we use (3.8) and (3.11) to estimate

$$\begin{aligned} &g'g^{iv} - 2g''g''' \\ &\quad < \frac{1}{\tilde{\epsilon}}\left(-g'(\eta)\int_0^\eta (g'')^2 e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds + 2g''(\eta)\int_0^\eta g'g''e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds\right), \end{aligned} \tag{3.12}$$

since $g'' > 0$ for $\eta \in (0, 1-\delta)$ and $g'''(0) > 0$. Let $\eta < \min\{\eta_3, \eta_1\}$, where η_1 is the zero point of g', which also approaches $\frac{1}{2}$ as $\tilde{\epsilon} \to 0$. Since $g' < 0$, $g'' > 0$, and $g''' > 0$,

$$\begin{aligned} g''(\eta)\int_0^\eta g'g''e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds &< \int_0^\eta g'(g'')^2 e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds \\ &< g'(\eta)\int_0^\eta (g'')^2 e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds. \end{aligned} \tag{3.13}$$

Applying (3.13) in (3.12), we have

$$g'g^{iv} - 2g''g''' < \frac{1}{\tilde{\epsilon}}g'(\eta)\int_0^\eta (g'')^2 e^{\frac{1}{\tilde{\epsilon}}\int_s^\eta gdr}ds < 0. \tag{3.14}$$

From (3.14) and (3.10), we see that $g^v < 0$ as long as $g''' > 0$ and $g' < 0$.

Since g''' is concave down and g'' is uniformly bounded on $[0, \frac{1}{2} - \sigma]$, the function g''' is uniformly bounded on $[0, \frac{1}{2} - \sigma]$, by Proposition 3.2, where $0 < \sigma \ll 1$. By Proposition 3.4, the second conclusion holds. $\square$

COROLLARY 3.6. *As* $\tilde{\epsilon} \to 0$, $g \sim h$ *uniformly on* $[0, 1-\delta]$ *for any given positive* $\delta \ll 1$.

Proof. For the given δ, $g(\eta)$ and $h(\eta)$ are nonzero on the interval $[\delta, 1-\delta]$; then, from the convergence of g on $[0, 1-\delta]$, we see that $g \sim h$ on $[\delta, 1-\delta]$, namely, $\lim \frac{g}{h} = 1$ uniformly on $[\delta, 1-\delta]$ as $\tilde{\epsilon} \to 0$. Since $g' - h' \to 0$ uniformly on $[0, \delta]$ (from Lemma 3.5), we see that $\lim \frac{g}{h} = 1$ for $\eta \in (0, \delta]$ uniformly. Therefore, $g \sim h$ on $[0, 1-\delta]$. $\square$

One of our goals in this section is to see how close g and h can be on $[0, 1-\delta]$. The result is given in Lemma 3.8, whose proof depends on Lemma 3.7.

LEMMA 3.7. *For any given positive* $\delta \ll 1$, $g''' - h''' \to 0$ *uniformly on* $[0, 1-\delta]$, *and* $g'' \sim h''$ *on* $[0, 1-\delta]$ *for sufficiently small* $\tilde{\epsilon} > 0$.

Proof. First, we prove that $g^{iv}(\eta)$ is uniformly bounded on $[0, 1-\delta]$. By Corollary 3.6, for a fixed positive $\sigma \ll 1$, the inequality

$$-\frac{1+\sigma}{\pi}\sin\pi\eta \le g(\eta) \le -\frac{1-\sigma}{\pi}\sin\pi\eta \tag{3.15}$$

holds for sufficiently small $\tilde{\epsilon} > 0$. From (3.15) and (3.8),

$$\begin{aligned} |g^{iv}| &\le \frac{\pi^2}{1-\Delta}\int_0^\eta \frac{(g'')^2}{\sin\pi s} d(e^{-\frac{1-\sigma}{\tilde{\epsilon}\pi}\int_s^\eta \sin\pi t dt}) \\ &= \frac{\pi^2}{1-\sigma}\frac{(g''(\xi))^2}{\sin\pi\xi}\left[1-e^{\frac{1-\sigma}{\tilde{\epsilon}\pi^2}(\cos\pi\eta-1)}\right] \\ &\le \frac{\pi^2}{1-\sigma}\frac{(g''(\xi))^2}{\sin\pi\xi}, \end{aligned} \tag{3.16}$$

where $\xi \in (0, \eta)$. Rewriting (3.16) as

$$|g^{iv}| \le \frac{\pi^2}{1-\sigma}g''(\xi)\frac{g''(\xi)}{\xi}\frac{\xi}{\sin\pi\xi},$$

one sees that g^{iv} is bounded because g'' and $g''(\xi)/\xi$ are bounded. Here, the boundedness of $g''(\xi)/\xi$ follows from that of g''' on $[0, 1-\delta]$. Then, the first conclusion of Lemma 3.7 follows from Proposition 3.4 since g, g', g'', g''', and g^{iv} are all uniformly bounded on $[0, 1-\delta]$.

To prove the second conclusion of Lemma 3.7, we use contradiction. If $\frac{g''}{h''}$ does not converge to 1 uniformly on $[0, 1-\delta]$, then there would be sequences $\{\tilde{\epsilon}_n\}$ and $\{\eta_n\}$ with $\tilde{\epsilon}_n \to 0$ and $\{\eta_n\} \subseteq [0, 1-\delta]$ such that $|\frac{g''(\eta_n)}{h''(\eta_n)} - 1| > \sigma$ for some constant $\sigma > 0$ and for all n. In addition, we can assume $\eta_n \to x \in [0, 1-\delta]$. Since $\delta > 0$ is fixed and $g''(1-\delta)h''(1-\delta) \ne 0$, the only possibility is $x = 0$. This is also impossible, for otherwise, it would imply the divergence of $(g''' - h''')$ at $\eta = 0$. □

An immediate consequence from Lemma 3.7 is that $g'''(0) = \tilde{\beta} \to \pi^2$ as $\tilde{\epsilon} \to 0$. This shows the relationship between $f'(0)$ and $f'''(0)$: $\frac{f'''(0)}{f'(0)} \to -\pi^2$ as $\tilde{\epsilon} \to 0$. With Lemma 3.7 and the expression of g^{iv} in (3.8), we can now prove the following lemma.

Lemma 3.8. *For any sufficiently small* $\delta > 0, g - h \to 0$ *in* C^4 *on* $[0, 1-\delta]$, *and* $g \sim h, g'' \sim h''$ *uniformly on* $[0, 1-\delta]$, *and* $g^{iv} \sim h^{iv}$ *uniformly on* $[\delta, 1-\delta]$.

Proof. Substituting the asymptotic formulas for g and g'' in (3.8), we obtain, by integration by parts,

(3.17)

$$g^{iv}(\eta) \sim -\left(\frac{\pi}{1-\Delta}\right)^3\left(\sin\frac{\pi\eta}{1-\Delta} + \frac{\pi}{1-\Delta}\int_0^\eta \cos\frac{\pi s}{1-\Delta}e^{\frac{(1-\sigma)^2}{\tilde{\epsilon}\pi^2}(\cos\frac{\pi\eta}{1-\Delta} - \cos\frac{\pi s}{1-\Delta})}ds\right)$$

uniformly on $[0, 1-\delta]$. Since

$$\int_0^{1-\delta}\left|\cos\frac{\pi s}{1-\Delta}\right|e^{\frac{(1-\sigma)^2}{\tilde{\epsilon}\pi^2}(\cos\frac{\pi\eta}{1-\Delta} - \cos\frac{\pi s}{1-\Delta})}ds \to 0$$

as $\tilde{\epsilon} \to 0$ by the dominated convergence theorem, we see that $g^{iv} \to h^{iv}$ uniformly on $[0, 1-\delta]$. In addition, it follows from (3.17) that $g^{iv} \sim h^{iv}$ uniformly on $[\delta, 1-\delta]$. □

Remark 1. It is false that $g^{iv} \sim h^{iv}$ uniformly on $[0, 1-\delta]$, for otherwise, if $\frac{g^{iv}(\eta,\tilde{\epsilon})}{h^{iv}(\eta,\tilde{\epsilon})} \to 1$ as $\tilde{\epsilon} \to 0$ uniformly on $[0, 1-\delta]$, then $\frac{g^{v}(0,\tilde{\epsilon})}{h^{v}(0,\epsilon)} \to 1$, a contradiction. In fact, for sufficiently small η the integral term in (3.18) is the leading term. Also, we cannot say that $g' \sim h', g''' \sim h'''$ on the interval because they do not vanish at the same points. Most importantly, we cannot have $g'' \sim h''$ at $\eta = 1-\Delta$ either, because $h'' = 0$ while $g'' \neq 0$ at $1-\Delta$.

Next, we extend the results of Lemma 3.8 from $[0, 1-\delta]$ to $[0, 1-\Delta]$.

The following proposition is a generalization of Proposition 3.4, which is applied to prove Theorem 3.1. Let $t_2 \equiv y_\epsilon$ denote the zero point of g'' in $(0,1)$. Then, $t_2 > 1-\Delta$, $g''(\eta) > 0$ for $\eta \in (0, t_2)$, and $t_2 \to 1$ as $\tilde{\epsilon} \to 0$, which is one of the previous results listed in section 1.

Proposition 3.9. *Suppose that* $\theta = \theta_{\tilde{\epsilon}} \in [1-\Delta, t_2]$ *and that* g, g', g'', g''',*and* g^{iv} *are all uniformly bounded on* $[0, \theta]$. *Then, as* $\tilde{\epsilon} \to 0$, *all* $g-h, g'-h', g''-h''$, *and* $g'''-h'''$ *converge to* 0 *uniformly on* $[0, \theta]$.

Remark 2. The uniform convergence on this "moving interval" is required for proving Lemma 6.1.

Proof. First, we see from (3.3) that $g'(\theta) \to 1$ since $g(\theta) \to 0$. Hence, $g''(\theta) \to 0$ from (3.1). By Lemmas 3.3, 3.5, 3.7, and 3.8, we see that only interval $[1-\delta, \theta]$ needs to be considered where $0 < \delta \ll 1$ is fixed.

Suppose, on the contrary, that the proposition is false for $g-h$. Then, there is a positive number σ and two sequences $\{\tilde{\epsilon}_n\}$ and $\{\eta_n\}, n = 1, 2, 3, ...$, such that $\eta_n \le \theta(n) (= \theta_{\tilde{\epsilon}_n}), |g(\eta_n, \tilde{\epsilon}_n) - h(\eta_n, \tilde{\epsilon}_n)| > \sigma$ for all n, and $\eta_n - \theta(n) \to 0$ as $n \to \infty$. Noting that $h(\theta) \to 0$ as $n \to \infty$ and $g \ge 0$ for $\eta \ge 1-\Delta$, we see that $g(\eta_n, \epsilon_n) > \frac{\delta}{2}$, and hence,

$$g'(\xi_n, \tilde{\epsilon}_n) = \frac{g(\eta_n, \tilde{\epsilon}_n) - g(\theta(n), \tilde{\epsilon}_n)}{\eta_n - \theta(n)} > \frac{\sigma}{3(\eta_n - \theta(n))},$$

which implies that g' is unbounded, a contradiction. The case $g''-h'' \to 0$ is handled the same way. To see $g'-h' \to 0$ uniformly on $[0, \theta]$, we assume, for contradiction, that $|g'(\eta_n, \tilde{\epsilon}_n) - h'(\eta_n, \tilde{\epsilon}_n)| > \sigma$ for $n = 1, 2, 3, ...$ and $\eta_n - \theta(n) \to 0$. Then, $g'(\eta_n)$ is either greater than $1 + \frac{\sigma}{2}$ or less than $1 - \frac{\sigma}{2}$. Choose $\delta = \frac{\sigma}{M}$ with $M > 4B$, where B is an upper bound for g'' on $[0, \theta]$. In the former case, there is a point $\xi_n \in (1-\delta, \eta_n)$ with $g''(\xi_n) > \frac{1+\sigma/2-g'(1-\delta)}{\eta_n-(1-\delta)} > \frac{\sigma}{2(\eta_n-1+\sigma/M)} > \frac{M}{3} > B$, a contradiction. Here, we have applied $g'(1-\delta) \approx h'(1-\delta) \approx 1-\delta$ by Lemma 3.8. The latter case is impossible, because $g'' > 0$. Similarly, $g'''-h''' \to 0$ uniformly on $[0, \theta]$, which implies $g'''(\theta) \to -\pi^2$. □

It is clear that we need to show the uniform boundedness of g, g', g'', g''', and g^{iv} on $[0, 1-\Delta]$ so that Proposition 3.9 can be applied.

Lemma 3.10. g, g', *and* g'' *are uniformly bounded on* $[0, 1-\Delta]$, $g^{(k)} - h^{(k)} \to 0$ *for* $k = 0, 1$, *and* $g(\eta) \sim -\frac{\sin[\pi\eta/(1-\Delta)]}{\pi/(1-\Delta)}$ *uniformly on* $[0, 1-\Delta]$ *as* $\tilde{\epsilon} \to 0$.

Proof. Recall that g is uniformly bounded and $g' > -\eta$ on $[0,1]$, and that $g'' > 0$ on $[0, 1-\Delta]$. Again, we assume a positive number $\delta \ll 1$ to be fixed. Since the convergence preserves the concavity of the solutions, $g''' < 0$ on $[\frac{1}{2}+\delta, 1-\delta]$, and hence, on $[\frac{1}{2}+\delta, 1]$. Then g' is concave down on $[\frac{1}{2}+\delta, 1-\Delta]$, which implies that g' must be bounded on the interval by Proposition 3.2. Furthermore, the concavity of g'' and the boundedness of g' on $[0, 1-\Delta]$ yield that g'' is bounded on $[0, 1-\Delta]$. The uniform convergence of g and g' follows from the proof of Proposition 3.9 because only the uniform boundedness of g'' is needed.

To prove the last conclusion of the lemma, we use contradiction again. Suppose it is false. Then, similar to above, there would be two sequences $\{\eta_n\}$ and $\{\tilde{\epsilon}_n\}$ such that as $n \to \infty$, $\eta_n - (1-\Delta) \to 0, \tilde{\epsilon}_n \to 0$, and $|\frac{g(\eta_n,\tilde{\epsilon}_n)}{h(\eta_n,\tilde{\epsilon}_n)} - 1| > \sigma$ for some $\sigma > 0$. Then,

$$\left|\frac{[g(\eta_n,\tilde{\epsilon}_n) - g(1-\Delta,\tilde{\epsilon}_n)]/(\eta_n - 1 + \Delta)}{[h(\eta_n,\tilde{\epsilon}_n) - h(1-\Delta,\tilde{\epsilon}_n)]/(\eta_n - 1 + \Delta)} - 1\right| > \sigma. \tag{3.18}$$

Applying the mean value theorem on both the numerator and denominator in (3.18), we see that for each n, there exist two numbers ξ_n and ζ_n such that

$$\left|\frac{g'(\xi_n,\tilde{\epsilon}_n)}{h'(\zeta_n,\tilde{\epsilon}_n)} - 1\right| > \sigma,$$

which implies either $g'(\xi_n,\tilde{\epsilon}_n) > 1 + \frac{\sigma}{2}$ or $g'(\xi_n,\tilde{\epsilon}_n) < 1 - \frac{\sigma}{2}$ since $h'(\zeta_n,\tilde{\epsilon}_n) \approx 1$. This would lead to g'' being unbounded, a contradiction. $\square$

By Lemma 3.7, $g'''(0) \to \pi^2$. It suffices to prove the uniform boundedness of g^{iv} on $[0, 1-\Delta]$ in order to get the uniform boundedness for g''' on $[0, 1-\Delta]$. Let $0 < \delta \ll 1$ be fixed. Since $h^v > 0$ on $[\frac{1}{2}+\delta, 1-\delta]$ and the convergence preserves the monotonicity, $g^v \geq 0$ on $[\frac{1}{2}+\delta, 1-\delta]$. From (3.6), we see $g^v(1-\Delta) < 0$ since $g''(1-\Delta) \neq 0$. It is then observed that g^v must have a zero in $[1-\delta, 1-\Delta)$. Let $t_5 = t_5(\tilde{\epsilon})$ denote the first zero of g^v in $[1-\delta, 1-\Delta)$. The structure of the function g^{iv} is given by the following lemma.

Lemma 3.11. *$g^v, g^{vi} \leq 0$ on $[t_5, 1-\Delta]$. Also, g''' and g^{iv} are uniformly bounded on $[0, 1-\Delta]$, and $g^{iv}(1-\Delta) \to 0$ as $\tilde{\epsilon} \to 0$.*

Proof. Since $g^v(1-\delta) \geq 0$ and t_5 is the first zero of g^v in $[1-\delta, 1-\Delta)$, we see that $g^{vi}(t_5) \leq 0$, and $g^{iv}(t_5) \geq g^{iv}(1-\delta)$ for sufficiently small $\delta > 0$. Hence, g^{iv} is uniformly bounded on $[0, t_5]$ because $g^{iv}(1-\delta)$ is uniformly bounded by Lemma 3.8. Therefore, g''', g'', and g' are all uniformly bounded on $[0, t_5]$. Next, we claim that $g^{vi}(\eta) \leq 0$ for $\eta \in [t_5, 1-\Delta]$. To see this, we assume that there is a first zero point t_6 of g^{vi} in $(t_5, 1-\Delta)$. Then, it suffices to prove $g^{vii}(t_6) < 0$, which is given as follows.

Differentiating (3.9) with respect to η once yields

$$\tilde{\epsilon} g^{vii} = g g^{vi} + 2g' g^v - g'' g^{iv} - 2(g''')^2. \tag{3.19}$$

From (3.9), we see at $\eta = t_6$, $g g^v + g' g^{iv} - 2g'' g''' = 0$, and therefore,

$$g^{iv} = \frac{2g'' g''' - g g^v}{g'}. \tag{3.20}$$

Substituting (3.20) into (3.19) leads to

$$\tilde{\epsilon} g^{vii} = \frac{[2(g')^2 + g g'']g^v - 2g'''[(g'')^2 + g' g''']}{g'} \tag{3.21}$$

at $\eta = t_6$. Since $g'' > 0$, $g'(1-\delta) \approx 1$, and $g'(1-\Delta) \to 1$, it must be that $g'(t_6) \approx 1$. Also, since $g''(1-\delta) \approx 0$, $g''' < 0$, and $g'' > 0$ for $\eta > 1-\delta$, we see $g''(t_6) \approx 0$. Thus, the first term in the numerator of (3.21) is negative because $g^v(t_6) < 0$, and the second term is approximately equal to $-2[g'''(t_6)]^2$ because $g'''(1-\delta) \approx -\pi^2(1-\delta)$ and $g^{iv} < 0$. This shows

$$\tilde{\epsilon} g^{vii}(t_6) < -[g'''(t_6)]^2 < 0.$$

Now, the picture is clear: $g^v > 0$ on $[\frac{1}{2}+\delta, t_5)$ and $g^v < 0, g^{vi} < 0$ on the interval $(t_5, 1-\Delta]$, in which g^{iv} reaches its local maximum at t_5 and then decreases. So, the bounds of $|g^{iv}|$ on $[1-\delta, 1-\Delta]$ are determined by the two ending points of the interval. By Lemma 3.8, $g^{iv}(1-\delta) \approx -\pi^3 \sin \pi(1-\delta)$. Since $t_5 > 1-\delta$, $g^{iv}(t_5) > g^{iv}(1-\delta)$, and $\delta > 0$ is arbitrarily chosen, it follows that $t_5 \to 1$ and $g^{iv}(t_5) \to 0$ as $\tilde{\epsilon} \to 0$. Therefore, to show the uniform boundedness of g^{iv} on $[0, 1-\Delta]$, we only need to prove the boundedness of $g^{iv}(1-\Delta)$. In fact, we can prove more: $g^{iv}(1-\Delta) \to 0$ as $\tilde{\epsilon} \to 0$. Suppose it is not so. Then, there exists a subsequence of $\{\tilde{\epsilon}\}$, denoted by $\{\tilde{\epsilon}_n\}, n = 1, 2, ...$, such that $g^{iv}(1-\Delta) < -\sigma$ for all $\tilde{\epsilon}_n$. From (3.1),

$$g^{iv}(1-\Delta) = \frac{-g'(1-\Delta)g''(1-\Delta)}{\tilde{\epsilon}_n}, \tag{3.22}$$

and from (3.6),

$$g^v(1-\Delta) = \frac{-[g''(1-\Delta)]^2}{\tilde{\epsilon}_n}. \tag{3.23}$$

Since $g^{vi} < 0$ and $g^{iv}(t_5) < 0$, we see from (3.22) that

$$g^v(1-\Delta) < \frac{g^{iv}(1-\Delta) - g^{iv}(t_5)}{1-\Delta-t_5} < \frac{-g'(1-\Delta)g''(1-\Delta)}{(1-\Delta-t_5)\tilde{\epsilon}_n}, \tag{3.24}$$

which, from (3.23), implies

$$\frac{-[g''(1-\Delta)]^2}{\tilde{\epsilon}_n} < \frac{-g'(1-\Delta)g''(1-\Delta)}{(1-\Delta-t_5)\tilde{\epsilon}_n}.$$

It turns out that

$$g''(1-\Delta) > \frac{g'(1-\Delta)}{1-\Delta-t_5},$$

which leads to $g''(1-\Delta)$ being unbounded, a contradiction of Lemma 3.10. The proof of Lemma 3.11 is complete. □

LEMMA 3.12. *As $\tilde{\epsilon} \to 0, g - h \to 0$ in C^4 and $g \sim h$ uniformly on $[0, 1-\Delta]$. In particular, at $\eta = 1-\Delta$, $g' \to 1, g'' \to 0, g''' \to -\pi^2$, and $g^{iv} \to 0$, as $\tilde{\epsilon} \to 0$.*

Proof. Let $1-\Delta = \theta$, and apply Lemmas 3.10 and 3.11 and Proposition 3.9. Then, $g \sim h$ uniformly on $[0, 1-\Delta]$, and the convergence holds up to C^3 on that interval. To prove the C^4 convergence, we apply the information about $g^{iv}(\eta)$ for $\eta \in [1-\delta, 1-\Delta]$ obtained in the proof of Lemma 3.11: the extreme values of g^{iv} on $[1-\delta, 1-\Delta]$ are determined at the two ending points. Since $g^{iv}(1-\delta) - h^{iv}(1-\delta) \to 0$ and $g^{iv}(1-\Delta) \to 0$, the uniform convergence of $g^{iv} - h^{iv}$ on $[1-\delta, 1-\Delta]$ follows. □

The proof of Theorem 3.1 is complete.

Our goal is to find the asymptotic behavior of the functions $f(\eta)$. It is seen from Theorem 3.1 that $f'(0)$ and Δ are to be determined. To do so, we first study the relation between these two quantities in the next section.

4. Relations among $\epsilon, \Delta_\epsilon$, and $f'(0, \epsilon)$.

To study the asymptotic behavior of the solution at the turning point, it is convenient to introduce a new independent variable

$$t \equiv \frac{\eta - z_\epsilon}{\Delta},$$

and define $f(\eta) = f(z_\epsilon + \Delta t) \equiv u(t)$ and $\epsilon^* \equiv \frac{\epsilon}{\Delta}$. Then, the turning point is $t = 0$, and $u = |\alpha| g$, $u' = |\alpha|\Delta g'$, $u'' = |\alpha|\Delta^2 g''$, $u''' = |\alpha|\Delta^3 g'''$, $u^{iv} = |\alpha|\Delta^4 g^{iv}$. The original equation (3.1), in terms of u and t, takes the form

$$\epsilon^* u^{iv} = uu''' - u'u''. \tag{4.1}$$

Lemma 3.12 applied to $u(t)$ at $t = 0$ implies that as $\epsilon \to 0, u - \frac{1-\Delta}{\pi} \sin \frac{\pi \Delta t}{1-\Delta} \to 0$ in $C^{iv}_{[-\frac{1-\Delta}{\Delta}, 0]}$, and

$$\frac{u'(0)}{|\alpha|\Delta} \to 1, \quad \frac{u''(0)}{|\alpha|\Delta^2} \to 0, \quad \frac{u'''(0)}{|\alpha|\Delta^3} \to -\pi^2, \quad \frac{u^{iv}(0)}{|\alpha|\Delta^4} \to 0. \tag{4.2}$$

The boundary conditions (1.3) now become

$$u(1) = 1, \qquad u'(1) = 0. \tag{4.3}$$

In addition, from the preliminary results listed in section 1, we see that $u^{iv}(t) < 0$ on $[-\frac{1-\Delta}{\Delta}, 1]$, and that $u(t) \geq 0, u'(t) \geq 0$ on $[0,1]$. Also, $u''(0) > 0$ for all $\epsilon^* > 0$. From (3.6), $u^v(t) < 0$ for $t \in [0,1]$. In this section we prove the following theorem.

THEOREM 4.1. *As $\epsilon \to 0$, the following conclusions hold:* (1) $\epsilon^* \to 0$, (2) $f'(0,\epsilon) \sim -\frac{1}{\Delta}$, *and* (3) $u(t, \epsilon^*) \sim t$ *uniformly on* $[0,1]$.

The proof of Theorem 4.1 consists of Lemmas 4.2, 4.3, and 4.4.

LEMMA 4.2. *There exists a positive number l such that $l \leq |\alpha|\Delta < 2$ for all $\epsilon > 0$.*

Proof. Since $u'''(0) < 0$ and $u^{iv} < 0$, we see $u''' < 0$ on $[0,1]$. An application of the concavity of $u'(t)$ for $t > 0$ shows $u'(t) \geq u'(0)(1-t)$ for $t \in [0,1]$. An integration of this inequality yields $u(1) \geq \frac{1}{2} u'(0)$. From (3.3), $g'(z_\epsilon) = \frac{f'(z_\epsilon)}{|\alpha|} = \frac{u'(0)}{|\alpha|\Delta} > \sqrt{1 + \tilde{\epsilon}\tilde{\beta}} > 1$ because $g'''(z_\epsilon) < 0$. Thus, $u'(0) > |\alpha|\Delta$, and hence, $|\alpha|\Delta < 2$. Since $u(t)$ is uniformly bounded on $[0,1]$ for all $\epsilon > 0$ and $u'(t) \geq 0$, $u'''(t) < 0$ on $[0,1]$, Proposition 3.2 can be applied. This concludes that u' is uniformly bounded above by a positive number $\frac{1}{l}$. The proof of Lemma 4.2 is complete. □

Since the behavior of the function $g(\eta)$ at the turning point has been found, we can now prove $\frac{\epsilon}{\Delta} \to 0$ as $\epsilon \to 0$, which was a crucial working assumption in [7], using (4.2) and Lemma 4.2.

LEMMA 4.3. $\epsilon^* \to 0$ *as* $\epsilon \to 0$.

Proof. Suppose not. Then, there would be a sequence $\{\epsilon_n^*\}, n = 1, 2, ...,$ such that either $\epsilon_n^* \to \infty$ or $\epsilon_n^* \to c > 0$ as $n \to \infty$. In the former case,

$$u^{iv} = \lambda_n (uu''' - u'u''), \tag{4.4}$$

where $\lambda_n = \frac{1}{\epsilon_n^*} \to 0$ as $n \to \infty$. Let $u_n \equiv u(t, \epsilon_n^*)$ be solutions of (4.4). By Lemma 4.2, one can assume that $u_n'(0) \to \gamma$ where $\gamma > 0$ is a constant. Then, $u_n''(0) = o(\Delta)$ and $u'''(0) = O\left(\Delta^2\right)$. Using the theorem that solutions continuously depend on initial conditions and parameters, we see $u_n(t) \approx \gamma t$ (since the limit equation $u^{iv} = 0$ with $u'(0) = \gamma, u(0) = u''(0) = u'''(0) = 0$ has the unique solution $u(t) = \gamma t$). This contradicts the boundary condition of $u(t)$ at $t = 1$. If $\epsilon_n^* \to \frac{1}{b} > 0$ and $u_n'(0) \to \mu \geq \frac{1}{l} > 0$ as $n \to \infty$, then $\lambda_n = \frac{1}{\epsilon_n^*} \to b > 0$. Since the limit equation

$$u^{iv} = b(uu''' - u'u'') \tag{4.5}$$

with the initial condition $u'(0) = \mu$, $u(0) = u''(0) = u'''(0) = 0$ has a unique solution $u(t) = \mu t$, we see $u_n(t) \approx \mu t$, again, a contradiction. Therefore, Lemma 4.3 holds. $\square$

Applying Lemma 4.3 and the information about $u(t)$ for $t \in [0, 1]$, we find the asymptotic behavior of $u(t)$ on $[0, 1-\delta]$ and the asymptotic relation between $\alpha \equiv f'(0)$ and Δ as follows.

Lemma 4.4. *As $\epsilon^* \to 0$, $|\alpha|\Delta \to 1$, $u(t) - t \to 0$ in $C^3_{[0,1-\delta]}$ for any given positive $\delta \ll 1$, and $u(t) \sim t$ uniformly on $[0, 1]$.*

Proof. From the proof of Lemma 4.2, $u'(t)$ is uniformly bounded on $[0, 1]$. Since $u''(0) = o(\Delta)$ and $u''' < 0$, we see that u'' is uniformly bounded on $[0, 1-\delta]$; otherwise u' would be unbounded on $[0, 1 - \frac{\delta}{2}]$. Similarly, u''' and u^{iv} are uniformly bounded on $[0, 1 - \delta]$ because $u^{iv}, u^{v} < 0$ for $t \in [0, 1]$. Then, applying an argument similar to the proof of Lemma 3.3 (mainly, an application of the Arzela–Ascoli theorem), we conclude that there exists a function $w(t)$ such that $u(t) \to w(t)$ in $C_{[0,1]}$ and in $C^3_{[0,1-\delta]}$ as $\epsilon^* \to 0$, and that $w(t)$ is neither the sine function nor the hyperbolic sine function because their fourth derivatives are positive. It turns out that $w(t) = t$. Then, the fact that $u(1) = 1$ forces $u'(0) \sim |\alpha|\Delta \to 1$. Furthermore, we claim that

$$\frac{u(t, \epsilon^*)}{t} \to 1$$

uniformly on $[0, 1]$. Suppose, on the contrary, that there exists a number $\sigma > 0$ and two sequences $\{\epsilon^*_n\}$ and $\{t_n\}$ with $\epsilon^*_n \to 0$ and $t_n \to 0$ as $n \to \infty$, such that $|\frac{u(t_n,\epsilon^*_n)}{t_n} - 1| > \sigma$ for all $n = 1, 2, 3, ...$ Then, by the mean value theorem, $|u'(\xi_n, \epsilon^*_n) - 1| > \sigma$ for all $n = 1, 2, 3, ...$ This contradicts the proved C^3 convergence. $\square$

The proof of Theorem 4.1 is complete.

By Lemma 4.4, $u'''(0) \sim -\Delta^2\pi^2$ as $\epsilon \to 0$. Therefore, we will be focusing on the approximation of $u'''(0)$ in order to find the asymptotic formula of Δ.

In the rest of the paper, the asymptotic behavior of the solution at the boundary layer $t = 1$ and at the turning point $t = 0$ is analyzed. The result is the following theorem.

Theorem 4.5. *As $\epsilon^* \to 0$,*

$$u''(1) \sim \frac{1}{\epsilon^*}, \quad u'''(1) \sim -\frac{1}{\epsilon^{*2}}, \quad u^{iv}(1) \sim -\frac{1}{\epsilon^{*3}}, \tag{4.6}$$

$$u'(0) = 1 + \epsilon^* + o(\epsilon^*), \tag{4.7}$$

$$u'''(0) \sim -(2e\pi)^{-\frac{1}{4}}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{4\epsilon^*}}, \tag{4.8}$$

and

$$u^{iv}(0) \sim -\pi\sqrt{e}\epsilon^{*-3}e^{-\frac{1}{2\epsilon^*}}, \qquad u''(0) \sim \pi\sqrt{e}\epsilon^{*-2}e^{-\frac{1}{2\epsilon^*}}, \tag{4.9}$$

where ϵ^* satisfies

$$\pi^2\Delta^2 \sim (2e\pi)^{-\frac{1}{4}}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{4\epsilon^*}}. \tag{4.10}$$

The proof of Theorem 4.5 is given in the remaining three sections. The first part of section 5 contains proofs of (4.6) and (4.7). Formulas (4.8), (4.9), and (4.10) are proved in the second part of section 5, and sections 6 and 7.

5. On $u(t)$ for $t \geq 0$. Differentiate (4.1) with respect to t once to get

$$\epsilon^* u^v = uu^{iv} - (u'')^2 .$$

Then, as in (3.7),

$$\left(u^{iv} e^{-\frac{1}{\epsilon^*}\int u(s)ds}\right)' = -\frac{1}{\epsilon^*}(u'')^2 e^{-\frac{1}{\epsilon^*}\int u(s)ds}, \tag{5.1}$$

which shows that the function $u^{iv}(t)e^{-\frac{1}{\epsilon^*}\int^t u(s)ds}$ is negative decreasing. Thus, for $t \in [0,1]$,

$$\left|u^{iv}(t)\right| \leq |u^{iv}(1)|e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}. \tag{5.2}$$

Therefore, to get the asymptotic behavior of $u(t)$ in a region including the right boundary layer, we need to estimate $u^{iv}(1)$.

LEMMA 5.1. *As $\epsilon^* \to 0$,*

$$u''(1) \sim -\frac{1}{\epsilon^*}, \quad u'''(1) \sim -\frac{1}{\epsilon^{*2}}, \quad u^{iv}(1) \sim -\frac{1}{\epsilon^{*3}}.$$

Proof. Successively integrating (4.1) yields

$$\epsilon^* u''' = uu'' - (u')^2 + [u'(0)]^2 + \epsilon^* u'''(0), \tag{5.3}$$

and

$$\epsilon^* u'' = uu' - 2\int_0^t (u')^2 ds + \{[u'(0)]^2 + \epsilon^* u'''(0)\}t + \epsilon^* u''(0). \tag{5.4}$$

Since $\lim_{\epsilon^* \to 0} u'(t) = 1$ pointwise in $[0,1)$ and $u'''(0) \sim -\pi^2\Delta^2 \to 0, u''(0) = o(\Delta)$ as $\epsilon^* \to 0$, we see from (5.4) that $\epsilon^* u''(1) + 1 \to 0$, by the dominated convergence theorem. This implies $u''(1) \sim -\frac{1}{\epsilon^*}$. It then follows from (5.3) and (4.1) that $u'''(1) \sim -\frac{1}{\epsilon^{*2}}$ and $u^{iv}(1) \sim -\frac{1}{\epsilon^{*3}}$ as $\epsilon^* \to 0$. $\square$

The asymptotic behavior of $u(t)$ for $t > 0$ in the outer region has been given by Lemma 4.4 already. The asymptotic formulas for u'' and u on the interval including the right boundary layer are presented in Lemma 5.3, whose proof requires Lemma 5.2.

LEMMA 5.2. *For sufficiently small $\epsilon^* > 0$,*

$$|u'''(0)| \leq 2e^{16}\epsilon^{*-\frac{7}{4}} e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds}.$$

Proof. Since $u^v < 0$ for $t > 0$, $u^{iv}(t) \leq u^{iv}(0)$ for $t \in [0,1]$. Thus, $u'''(t) \leq u^{iv}(0)\,t + u'''(0)$ and $u''(t) \leq \frac{1}{2}u^{iv}(0)t^2 + u'''(0)t + u''(0)$ for $t \in [0,1]$. Let $a > 0$ be the zero of u''. It follows from the last inequality that

$$\frac{1}{2}u^{iv}(0)a^2 + u'''(0)a + u''(0) \geq 0. \tag{5.5}$$

Note from (4.1) that $u''(0) = \frac{-\epsilon^* u^{iv}(0)}{u'(0)}$, and substitute this equation into (5.5) to get

$$\frac{1}{2}u^{iv}(0)u'(0)a^2 - \epsilon^* u^{iv}(0) \geq \frac{1}{2}u^{iv}(0)u'(0)a^2 + u'(0)u'''(0)a - \epsilon^* u^{iv}(0) \geq 0.$$

Then, $a^2 < \frac{2\epsilon^*}{u'(0)}$, and hence, $a < 2\sqrt{\epsilon^*}$ for sufficiently small ϵ^* since $u'(0) \to 1$ as $\epsilon^* \to 0$. Thus $u'', u''', u^{iv} < 0$ for $t > a$. Since u'' is concave down, $u''(t) < u'''(a)(t-a)$ for $t \geq a$. On the other hand, integrating (5.1) and solving the resulting equation produces for any $t \in [-\frac{1-\Delta}{\Delta}, 1]$

$$u^{iv}(t) = -\frac{1}{\epsilon^*}\int_0^t [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^t u(r)dr} + u^{iv}(0)e^{\frac{1}{\epsilon^*}\int_0^t u(r)dr}. \tag{5.6}$$

Since $u^{iv}(0) < 0$,

$$\begin{aligned} |u^{iv}(t)| &> \frac{1}{\epsilon^*}e^{\frac{1}{\epsilon^*}\int_0^t u(s)ds}\int_a^t (u''(s))^2 e^{-\frac{1}{\epsilon^*}\int_0^s u(r)dr}ds \\ &\geq \frac{1}{\epsilon^*}e^{\frac{1}{\epsilon^*}\int_0^t u(s)ds}[u'''(a)]^2\int_{a+\delta_1}^{a+\delta_2}(s-a)^2 e^{-\frac{1}{\epsilon^*}\int_0^s u(r)dr}ds \\ &\geq \frac{1}{\epsilon^*}e^{\frac{1}{\epsilon^*}\int_0^t u(s)ds}[u'''(a)]^2(\delta_2-\delta_1)\delta_1^2 e^{-\frac{1}{\epsilon^*}\int_0^{a+\delta_2} u(r)dr} \end{aligned} \tag{5.7}$$

for any two δ_1 and δ_2 with $t > \delta_2 > \delta_1 > 0$, which implies

$$|u'''(a)|^2 \leq \frac{\epsilon^*|u^{iv}(t)|}{(\delta_2-\delta_1)\delta_1^2}e^{-\frac{1}{\epsilon^*}\int_{a+\delta_2}^t u(s)ds}. \tag{5.8}$$

Setting $t = 1, \delta_1 = \sqrt{\epsilon^*}, \delta_2 = 2\delta_1$ in (5.8), we have, for sufficiently small $\epsilon^* > 0$,

$$|u'''(a)|^2 \leq e^{-\frac{1}{\epsilon^*}\int_0^1 u(s)ds}\frac{|u^{iv}(1)|}{\sqrt{\epsilon^*}}e^{\frac{1}{\epsilon^*}\int_0^{a+2\sqrt{\epsilon^*}} u(s)ds}. \tag{5.9}$$

Since $a < 2\sqrt{\epsilon^*}$, $u(s) \leq 2s$, and $|u^{iv}(1)| \leq \frac{2}{\epsilon^*}$ for sufficiently small $\epsilon^* > 0$, we see from (5.9) that

$$|u'''(0)| \leq 2e^{16}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds}. \tag{5.10}$$

This completes the proof of the lemma. □

LEMMA 5.3. *For any given* $0 < \delta \ll 1$, *as* $\epsilon^* \to 0$,

$$u(t) \sim u(\delta) + u'(\delta)(t-\delta) + \frac{\epsilon^{*4}u^{iv}(1)}{u^4(t)}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}, \tag{5.11}$$

$$u''(t) \sim u''(\delta) + u'''(\delta)(t-\delta) + \frac{\epsilon^{*2}u^{iv}(1)}{u^2(t)}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \tag{5.12}$$

uniformly on $[\delta, 1]$.

Proof. Since $|u^{iv}(1)| \leq 2\epsilon^{*-3}$, we see from (5.2) that

$$|u^{iv}(t)| \leq 2{\epsilon^*}^{-3}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \tag{5.13}$$

for all $t \in [0, 1]$. Integrating (5.13) from 0 to t and applying Lemma 5.2, we find that

$$\begin{aligned} |u'''(t)| &\leq |u'''(0)| + 2{\epsilon^*}^{-3}\int_0^{\frac{1}{4}} e^{-\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx \\ &\leq |u'''(0)| + 2{\epsilon^*}^{-3}e^{-\frac{1}{\epsilon^*}\int_{\frac{1}{4}}^1 u(s)ds} \\ &\leq Q\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} \end{aligned} \tag{5.14}$$

for $t \le \frac{1}{4}$, where $Q = 2e^{16} + 1$, and that

$$\begin{aligned} |u'''(t)| &= |u'''(0)| + 4\epsilon^{*-3}\left(\int_0^{\frac{1}{4}} + \int_{\frac{1}{4}}^{t}\right) e^{-\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx \\ &\le Q_1\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} + 4\epsilon^{*-3}\int_{\frac{1}{4}}^{t} e^{-\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx \\ &\le Q_1\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} + N_1\epsilon^{*-2}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \qquad (5.15) \end{aligned}$$

for $t > \frac{1}{4}$, where $N_1 > 0$ is a constant and $Q_1 = 1+Q$. Here, the method of integration by parts has been utilized when we estimate the leading term of $\int_{\frac{1}{4}}^{t} e^{-\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx$. Similarly, integrating (5.14) and (5.15) from 0 to t, respectively, we find that

$$|u''(t)| \le |u''(0)| + Q\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds}$$

for $t \le \frac{1}{4}$, and

$$|u''(t)| \le |u''(0)| + Q_1\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} + N_2\epsilon^{*-1}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \qquad (5.16)$$

for $t > \frac{1}{4}$, where $N_2 > 0$ is a constant. Since $u''(0) = -\epsilon^* u^{iv}(0)$, we see from (5.2) and Lemma 5.1 that

$$\begin{aligned} |u''(t)| &\le 2{\epsilon^*}^{-2}e^{-\frac{1}{\epsilon^*}\int_0^1 u(s)ds} + Q\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} \\ &< M_1\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} \qquad (5.17) \end{aligned}$$

for $t \le \frac{1}{4}$, and

$$|u''(t)| \le Q_1\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{2\epsilon^*}\int_0^1 u(s)ds} + N_2\epsilon^{*-1}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \qquad (5.18)$$

for $t > \frac{1}{4}$. Thus, for all $t \in [0,1]$, the inequality (5.18) holds.

An integration of (5.1) from t to 1 gives

$$u^{iv}(1) - u^{iv}(t)e^{-\frac{1}{\epsilon^*}\int_1^t u(s)ds} = -\frac{1}{\epsilon^*}\int_t^1 (u'')^2 e^{-\frac{1}{\epsilon^*}\int_1^t u(s)ds}dt,$$

which implies

$$u^{iv}(t) = \left(u^{iv}(1) - \frac{1}{\epsilon^*}\int_t^1 (u'')^2 e^{\frac{1}{\epsilon^*}\int_r^1 u(s)ds}dr\right)e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}. \qquad (5.19)$$

Substituting (5.18) into (5.19), we obtain

$$\begin{aligned} |u^{iv}(t) - u^{iv}(1)e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}| &\le \frac{2}{\epsilon^*}\Bigg(\int_t^1 [Q_1^2\epsilon^{*-\frac{7}{2}}e^{-\frac{1}{\epsilon^*}\int_0^1 u(s)ds} \\ &\quad + N_2^2\epsilon^{*-2}e^{-\frac{2}{\epsilon^*}\int_x^1 u(s)ds}]e^{\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx\Bigg)e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \\ &\le 2Q_1^2\epsilon^{*-\frac{9}{2}}e^{-\frac{1}{\epsilon^*}\int_0^1 u(s)ds} \qquad (5.20) \\ &\quad + 2N_2^2\epsilon^{*-3}e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}\int_t^1 e^{-\frac{1}{\epsilon^*}\int_x^1 u(s)ds}dx. \end{aligned}$$

Again, using integration by parts on the second term on the right-hand side of (5.20), we find that for any $t \in [\delta, 1]$,

$$\left| u^{iv}(t) - u^{iv}(1) e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} \right| \leq 2Q_1^2 \epsilon^{*-\frac{9}{2}} e^{-\frac{1}{\epsilon^*}\int_0^1 u(s)ds} + N_3 \epsilon^{*-2} e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}, \tag{5.21}$$

where N_3 is a constant depending only on δ. An integration of (5.21) over (δ, t), where $0 < \delta \ll 1$ is fixed, shows that for $t \in [\delta, 1]$,

$$u'''(t) = u'''(\delta) + \frac{\epsilon^* u^{iv}(1)}{u(t)} e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} (1 + O(\epsilon^*)) + R(t, \epsilon^*), \tag{5.22}$$

where $R(t, \epsilon^*)$ is an exponentially small term. Successively integrating (5.22) over $[\delta, t]$ yields

$$\begin{aligned} u''(t) = {} & u''(\delta) + u'''(\delta)(t - \delta) \\ & + \frac{\epsilon^{*2} u^{iv}(1)}{u^2(t)} e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} (1 + O(\epsilon^*)) + R_2(t, \epsilon^*), \end{aligned} \tag{5.23}$$

and

$$\begin{aligned} u(t) = {} & u(\delta) + u'(\delta)(t - \delta) + u''(\delta)\frac{(t-\delta)^2}{2} + \frac{u'''(\delta)(t-\delta)^3}{3} \\ & + \frac{\epsilon^{*4} u^{iv}(1)}{u^4(t)} e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} (1 + O(\epsilon^*)) + R_0(t, \epsilon^*), \end{aligned} \tag{5.24}$$

where R_0 and R_2 are exponentially small terms. Since $u''(\delta)$ and $u'''(\delta)$ are of smaller order than that of $e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds}$ for the fixed $\delta < \frac{1}{4}$, from (5.17) and (5.14), Lemma 5.3 follows immediately. $\square$

COROLLARY 5.4. *For sufficiently small* $\epsilon^* > 0$,

$$u'(0) = 1 + \epsilon^* + o(\epsilon^*)$$

and

$$\frac{1}{\epsilon^*}\int_0^1 u(s)ds = \frac{1}{2\epsilon^*} + \frac{1}{2} + O(\epsilon^*).$$

Proof. Applying Lemma 5.3 for $t = 1$, one finds from (5.11) that for a fixed positive $\delta \ll 1$,

$$1 \sim u(\delta) + u'(\delta)(1 - \delta) + \epsilon^{*4} u^{iv}(1). \tag{5.25}$$

Since $u'(\delta) = u'(0) + u''(\xi_1)\delta$ and $u(\delta) = u'(0)\delta + u''(\xi_2)\frac{\delta^2}{2}$, where $\xi_i \in (0, \delta)$ and $u''(\xi_i)$ is exponentially small for $i = 1, 2$ from (5.17), we see that

$$1 \sim u'(0) + \epsilon^{*4} u^{iv}(1),$$

which implies the first conclusion of the corollary. To obtain the second conclusion, we integrate (5.11) over $[\delta, t]$. Then,

$$\int_\delta^t u(s)ds \sim u(\delta)\,(t - \delta) + u'(\delta)\frac{(t-\delta)^2}{2} + \frac{\epsilon^5 u^{iv}(1)}{u^5(t)} e^{-\frac{1}{\epsilon^*}\int_t^1 u(s)ds} (1 + O(\epsilon^*)).$$

Since $u'(\delta) \sim u'(0)$ and $u(s) \sim u'(0)s$ for $s \in [0, \delta]$,

$$\int_0^t u(s)ds = \left(\int_0^\delta + \int_\delta^t\right) uds \sim \frac{1}{2}\{[u'(0)\delta]^2 + [2u'(0)\delta(t-\delta) + u'(0)(t-\delta)]^2\} + O(\epsilon^{*2})$$

for any $t > \delta$. Thus, for all $t \in [\delta, 1]$,

$$\int_0^t u(s)ds \sim \frac{1}{2}(1+\epsilon^*)t^2 + O(\epsilon^{*2}) \tag{5.26}$$

and

$$\frac{1}{\epsilon^*}\int_0^1 u(s)ds \sim \frac{1}{2\epsilon^*} + \frac{1}{2} + O(\epsilon^*). \tag{5.27}$$

In addition, if $0 \le t \le \delta$, then $u'(t) \sim 1+\epsilon^*$, and hence, $u(t) \sim (1+\epsilon^*)t$. Therefore, (5.26) still holds for $t \in [0, \delta)$. □

The investigation of the asymptotic behavior of $u(t)$ on $(0, 1]$, and, in particular, at the right boundary layer, is complete. This enables us to prove the first asymptotic formula linking $u^{iv}(0)$ and $u'''(0)$ given in the following lemma.

LEMMA 5.5. As $\epsilon^* \to 0$, $u^{iv}(0) \sim [u'''(0)]^2\sqrt{\frac{\pi\epsilon^*}{2}} - \epsilon^{*-3}e^{-(\frac{1}{2}+\frac{1}{2\epsilon^*})}$.

Proof. Evaluating (5.6) at $t = 1$, we obtain

$$u^{iv}(1) = -\frac{1}{\epsilon^*}\int_0^t [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} + u^{iv}(0)e^{\frac{1}{\epsilon^*}\int_0^1 u(r)dr}. \tag{5.28}$$

From (5.28) and (5.27), and by Lemma 5.1,

$$-\frac{1}{\epsilon^{*3}} \sim -\frac{1}{\epsilon^*}\int_0^\delta [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} - \frac{1}{\epsilon^*}\int_\delta^1 [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} + u^{iv}(0)e^{\frac{1}{2\epsilon^*}+\frac{1}{2}+O(\epsilon^*)}, \tag{5.29}$$

where $0 < \delta \ll 1$ is fixed. Substituting (5.23) into the second term on the right-hand side of (5.29), we find

$$\begin{aligned}\frac{1}{\epsilon^*}\int_\delta^1 [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} &\sim \frac{1}{\epsilon^*}\int_\delta^1 \{u''(\delta) + u'''(\delta)(s-\delta)\}^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \\ &+ \frac{1}{\epsilon^*}\int_\delta^1 \frac{\epsilon^{*4}(u^{iv}(1))^2}{u^4(s)} e^{-\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds,\end{aligned} \tag{5.30}$$

where the cross product terms in the expansion of $[u(s)]^2$ disappeared because they are negligible, for example, for the fixed δ,

$$\frac{1}{\epsilon^*}\int_\delta^1 \frac{\epsilon^{*2}u^{iv}(1)}{u^2(s)} e^{-\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \sim -\frac{1}{\epsilon^{*2}}\int_\delta^1 \frac{1}{u^2(s)} e^{-\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds = O(\epsilon^{*-2}).$$

Similarly, the last term on the right-hand side of (5.30) is negligible. Thus,

$$\begin{aligned}-\frac{1}{\epsilon^{*3}} \sim &- \frac{1}{\epsilon^*}\int_0^\delta [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \\ &- \frac{1}{\epsilon^*}\int_\delta^1 [u''(\delta) + u'''(\delta)(s-\delta)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \\ &+ u^{iv}(0)e^{\frac{1}{2\epsilon^*}+\frac{1}{2}+O(\epsilon^*)}.\end{aligned} \tag{5.31}$$

Apply Taylor's theorem to get $u''(s) = u''(0) + u'''(0)s + u^{iv}(x)\frac{s^2}{2}$ where $x \in (0, s)$. From (5.13), and since $u''(0) \sim \epsilon^* u^{iv}(0)$, the first integral in (5.31) admits

$$-\frac{1}{\epsilon^*}\int_0^\delta [u''(s)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \sim -\frac{1}{\epsilon^*}[u'''(0)]^2 \int_0^\delta s^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds.$$

Here, the terms having the fourth derivative $u^{iv}(x)$ are neglected because they are of smaller order than ϵ^{*-3}. Similarly, since $u''(\delta) = u''(0) + u'''(0)\delta + u^{iv}(b)\delta^2/2$ and $u'''(\delta) = u'''(0) + u^{iv}(c)\delta$ where b, $c \in (0, \delta)$,

$$\begin{aligned}&\frac{1}{\epsilon^*}\int_\delta^1 [u''(\delta) + u'''(\delta)(s-\delta)]^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds \\ &\sim \frac{1}{\epsilon^*}[u'''(0)]^2 \int_\delta^1 s^2 e^{\frac{1}{\epsilon^*}\int_s^1 u(r)dr} ds.\end{aligned}$$

Finally,

$$\begin{aligned}-\frac{1}{\epsilon^{*3}} &\sim -\frac{[u'''(0)]^2}{\epsilon^*} e^{\frac{1}{\epsilon^*}\int_0^1 u(r)dr} \int_0^1 s^2 e^{-\frac{1}{\epsilon^*}\int_0^s u(r)dr} ds + u^{iv}(0) e^{\frac{1}{2\epsilon^*}+\frac{1}{2}+O(\epsilon^*)} \\ &\sim -\frac{[u'''(0)]^2}{\epsilon^*} e^{\frac{1}{2\epsilon^*}+\frac{1}{2}} \int_0^1 s^2 e^{-(\frac{1}{2\epsilon^*}+\frac{1}{2})s^2} ds + u^{iv}(0) e^{\frac{1}{2\epsilon^*}+\frac{1}{2}}. \qquad (5.32)\end{aligned}$$

Make a substitution $x = (\frac{1}{2\epsilon^*} + \frac{1}{2})s^2$ in the integral of (5.32). Then,

$$\int_0^1 s^2 e^{-(\frac{1}{2\epsilon^*}+\frac{1}{2})s^2} ds = \frac{1}{2}\left(\frac{2\epsilon^*}{1+\epsilon^*}\right)^{\frac{3}{2}} \int_0^{\frac{1+\epsilon^*}{2\epsilon^*}} x^{\frac{1}{2}} e^{-x} dx \sim \sqrt{2}\left(\frac{\epsilon^*}{1+\epsilon^*}\right)^{3/2} \Gamma\left(\frac{3}{2}\right),$$

and hence,

$$-\frac{1}{\epsilon^{*3}} \sim -\left\{\epsilon^{*-1}\sqrt{2}\left(\frac{\epsilon^*}{1+\epsilon^*}\right)^{3/2}\Gamma\left(\frac{3}{2}\right)[u'''(0)]^2 - u^{iv}(0)\right\} e^{\frac{1}{2}+\frac{1}{2\epsilon^*}},$$

from which the lemma follows. □

Our goal is to determine the asymptotic value of $u'''(0)$. Therefore, another asymptotic formula linking $u'''(0)$ and $u^{iv}(0)$ is needed. Since Lemma 5.5 is obtained using only information about $u(t)$ for $t \geq 0$, we must return to the left side of the turning point in section 6 to find how these two quantities are related for $t \leq 0$.

6. On $u(t)$ for $t \leq 0$. For convenience, we rewrite Lemma 3.12 as the following in terms of the function $u(t)$.

LEMMA 6.1. $\Delta u(t) - (-\frac{1-\Delta}{\pi}\sin\frac{\pi\Delta t}{1-\Delta}) \to 0$ in C^4 and $|u^{(k)}(t)| = O(\Delta^{k-1})$ for $k = 1, 2, 3, 4$ on $[-\frac{1-\Delta}{\Delta}, 0]$, and $u(t) \sim -\frac{1-\Delta}{\Delta\pi}\sin\frac{\pi\Delta t}{1-\Delta}$ uniformly on $[-\frac{1-\Delta}{\Delta}, 0]$.

From (5.6),

$$u^{iv}(0) = u^{iv}(-M)e^{-\frac{1}{\epsilon^*}\int_0^{-M} u(r)dr} + \frac{1}{\epsilon^*}\int_0^{-M} [u''(s)]^2 e^{-\frac{1}{\epsilon^*}\int_0^s u(r)dr} ds \qquad (6.1)$$

for any $M \in [0, \frac{1-\Delta}{\Delta}]$. To determine the leading term of $u^{iv}(0)$, we must find an asymptotic formula for the integrals of $u(r)$ in (6.1).

Lemma 6.2. *As* $\epsilon^* \to 0$,

$$\int_0^r u(t)dt \sim 2\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2\frac{\pi\Delta r}{2(1-\Delta)}, \tag{6.2}$$

and

$$\int_0^r [u(s)]^2 e^{-\frac{1}{\epsilon^*}\int_0^s u(t)dt} \sim \sqrt{\frac{\pi}{2}}\epsilon^{*\frac{3}{2}}[u'''(0)]^2 \tag{6.3}$$

uniformly for all $r \in [-\frac{1-\Delta}{\Delta}, -\rho]$ *where* $\rho > 0$ *is a constant.*

Proof. Recall from section 1 that $g(\eta) \sim h(\eta) = -\frac{1-\Delta}{\pi}\sin\frac{\pi\eta}{1-\Delta}$ uniformly on $[0, 1-\Delta]$ as $\tilde{\epsilon} \to 0$. Hence,

$$\int_{1-\Delta}^{\tau} g(\eta)d\eta \sim \int_{1-\Delta}^{\tau} h(\eta)d\eta = \left(\frac{1-\Delta}{\pi}\right)^2\left(1+\cos\frac{\pi\tau}{1-\Delta}\right) \tag{6.4}$$

uniformly for $\tau \in [0, 1-\Delta]$. Let $\eta = 1-\Delta+\Delta t$ and $r = (\tau - 1 + \Delta)/\Delta$. If follows that

$$\frac{\Delta}{|\alpha|}\int_0^r u(t)dt \sim 2\left(\frac{1-\Delta}{\pi}\right)^2 \sin^2\frac{\pi\Delta r}{2(1-\Delta)},$$

which implies (6.2) by Lemma 4.4. To prove the asymptotic formula (6.3), we directly apply the definition. For any given $\sigma > 0$, from (6.2), there exists an ϵ_0^* such that if $\epsilon^* < \epsilon_0^*$, then for all $r \in [-\frac{1-\Delta}{\Delta}, 0]$,

$$e^{-\frac{2(1+\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta r}{2(1-\Delta)}} \le e^{-\frac{1}{\epsilon^*}\int_0^r u(t)dt} \le e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta r}{2(1-\Delta)}}, \tag{6.5}$$

and

$$\begin{aligned}\int_0^{|r|}(u'')^2 e^{-\frac{2(1+\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds &\le \int_0^{|r|}(u'')^2 e^{-\int_0^s\frac{u(t)}{\epsilon^*}dt}ds\\ &\le \int_0^{|r|}(u'')^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds. \end{aligned}\tag{6.6}$$

Set an $r \ge \rho$. A substitution of the expression $u''(s) = u''(0) + u'''(0)s + u^{iv}(\xi)s^2/2$, where $\xi \in (0, s)$ depending on s, into the right integral of (6.6) leads to

$$\begin{aligned}&\int_0^{|r|}(u'')^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds\\ &\quad\sim [u'''(0)]^2\int_0^{|r|}s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds,\end{aligned}\tag{6.7}$$

for all $\epsilon^* < \epsilon_1^*$, where $\epsilon_1^* \in (0, \epsilon_0^*)$ depending only on ρ and σ. The terms including $u''(0)$ and $u^{iv}(z)$ in the resulting expansion are neglected because $u^{iv}(t)| = \Delta^3|g^{iv}| = O(\Delta^3) \ll |u'''(0)|$, $u''(0) \sim -\epsilon^* u^{iv}(0) = o(\Delta^3) \ll |u'''(0)|$. To see that ϵ_1^* is independent of r, we rewrite the integral on the right-hand side of (6.7) as

$$\begin{aligned}&\int_0^{|r|}s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds\\ &\quad= \left\{\int_0^{\rho} + \int_{\rho}^{|r|}\right\} s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2\sin^2\frac{\pi\Delta s}{2(1-\Delta)}}ds.\end{aligned}\tag{6.8}$$

Since $\frac{2}{\pi} \le \sin\theta \le \theta$ for all $\theta \in (0, \frac{\pi}{2})$,

$$\int_\rho^{|r|} s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2 \frac{\pi\Delta s}{2(1-\Delta)}} ds \le \int_\rho^{|r|} s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*\pi^2}s^2} ds = M_1 {\epsilon^*}^{\frac{3}{2}} \int_{k_2(\rho,\epsilon^*)}^{k_1(r,\epsilon^*)} x^{\frac{1}{2}} e^{-x} dx,$$

where $k_1(\epsilon^*) = \frac{2(1-\sigma)}{\epsilon^*\pi^2} r^2 \to \infty$, $k_2(\epsilon^*) = \frac{2(1-\sigma)}{\epsilon^*\pi^2}\rho^2 \to \infty$, and $M_1 = 2^{-\frac{5}{2}}\pi^3(1-\sigma)^{-\frac{3}{2}}$. This shows that the second integral on the right-hand side of (6.8) is of order $o({\epsilon^*}^{\frac{3}{2}})$ for all $r \in [-\frac{1-\Delta}{\Delta}, -\rho]$ as $\epsilon^* \to 0$. On the other hand, noting from Lemma 5.2 that Δ is exponentially small as $\epsilon^* \to 0$, we see that

$$\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2 \frac{\pi\Delta s}{2(1-\Delta)} \sim \frac{(1-\sigma)\, s^2}{2\epsilon^*}$$

uniformly on $[0, \rho]$, and

$$\begin{aligned} \int_0^\rho s^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2 \frac{\pi\Delta s}{2(1-\Delta)}} ds &\sim \int_0^\rho s^2 e^{-\frac{(1-\sigma)s^2}{2\epsilon^*}} ds \\ &= \sqrt{2}\left(\frac{\epsilon^*}{1-\sigma}\right)^{\frac{3}{2}} \int_0^{\frac{\rho^2(1-\sigma)}{2\epsilon^*}} x^{\frac{1}{2}} e^{-x} dx \\ &\sim \sqrt{2}\left(\frac{\epsilon^*}{1-\sigma}\right)^{\frac{3}{2}} \Gamma\left(\frac{3}{2}\right), \end{aligned} \tag{6.9}$$

which implies that the first integral on the right-hand side of (6.8) contributes the leading term for (6.7). This also proves

$$\int_0^{|r|} [u''(s)]^2 e^{-\frac{2(1-\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2 \frac{\pi\Delta s}{2(1-\Delta)}} ds \sim \sqrt{2}\left(\frac{\epsilon^*}{1-\sigma}\right)^{\frac{3}{2}} \Gamma\left(\frac{3}{2}\right). \tag{6.10}$$

Similarly,

$$\int_0^{|r|} [u''(s)]^2 e^{-\frac{2(1+\sigma)}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2 \sin^2 \frac{\pi\Delta s}{2(1-\Delta)}} ds \sim \sqrt{2}\left(\frac{\epsilon^*}{1+\sigma}\right)^{\frac{3}{2}} \Gamma\left(\frac{3}{2}\right). \tag{6.11}$$

From (6.6) and these last two formulas, we conclude that for the given $\sigma > 0$, there is an $\epsilon_1^* > 0$ such that

$$\sqrt{2}\left(\frac{\epsilon^*}{1+2\sigma}\right)^{\frac{3}{2}} \Gamma\left(\frac{3}{2}\right) < \int_0^{|r|} (u'')^2 e^{-\int_0^s \frac{u(t)}{\epsilon^*} dt} ds < \sqrt{2}\left(\frac{\epsilon^*}{1-2\sigma}\right)^{\frac{3}{2}} \Gamma\left(\frac{3}{2}\right).$$

Therefore, (6.3) holds, and the proof of Lemma 6.2 is complete. □

We can now prove the second asymptotic relation between $u'''(0)$ and $u^{iv}(0)$ as follows.

LEMMA 6.3. *As* $\epsilon^* \to 0, u^{iv}(0) \sim -\sqrt{\frac{\pi\epsilon^*}{2}}[u'''(0)]^2$.

Proof. Set $M = \frac{1-\Delta}{\Delta}$ in (6.1). By Lemma 6.2,

$$u^{iv}(0) \sim u^{iv}(-M) e^{-\frac{1}{\epsilon^*}\int_0^{-M} u(r)dr} - \frac{[u'''(0)]^2}{\epsilon^*}\sqrt{2}{\epsilon^*}^{\frac{3}{2}}\Gamma\left(\frac{3}{2}\right). \tag{6.12}$$

From (6.5), we see that for $\epsilon^* < \epsilon_1^*$, where ϵ_1^* is given in the proof of Lemma 6.2,

$$|u^{iv}(-M)| e^{-\frac{1}{\epsilon^*}\int_0^{-M} u(r)dr} \le |u^{iv}(-M)| e^{-\frac{1-\sigma}{\epsilon^*}\left(\frac{1-\Delta}{\Delta\pi}\right)^2}.$$

Since $u^{iv}(-M) = O(\Delta^3), u'''(0) \sim \pi^2\Delta^2$, and Δ is exponentially small, the second term on the right-hand side of (6.12) provides the leading term for $u^{iv}(0)$, i.e.,

$$u^{iv}(0) \sim -\frac{[u'''(0)]^2}{\epsilon^*}\sqrt{2}\epsilon^{*\frac{3}{2}}\Gamma\left(\frac{3}{2}\right). \tag{6.13}$$

The proof of Lemma 6.3 is complete. □

7. At the turning point. The two asymptotic formulas linking $u'''(0)$ and $u^{iv}(0)$ have been determined in sections 5 and 6. In this final section, we combine them to determine the asymptotic values of these two quantities. The asymptotic value of Δ_ϵ is then determined in terms of the original parameter ϵ, which will complete the proof of the main result of the paper.

LEMMA 7.1. *As* $\epsilon^* \to 0$,

$$u'''(0) \sim -(2e\pi)^{-\frac{1}{4}}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{4\epsilon^*}}$$

and

$$u^{iv}(0) \sim -\pi\sqrt{e}\epsilon^{*-3}e^{-\frac{1}{2\epsilon^*}}, \qquad u''(0) \sim \pi\sqrt{e}\epsilon^{*-2}e^{-\frac{1}{2\epsilon^*}}.$$

Proof. From Lemmas 5.5 and 6.3,

$$-[u'''(0)]^2\sqrt{\frac{\pi\epsilon^*}{2}} \sim [u'''(0)]^2\sqrt{\frac{\pi\epsilon^*}{2}} - \epsilon^{*-3}e^{-(\frac{1}{2}+\frac{1}{2\epsilon^*})}.$$

Then

$$u'''(0) \sim -(2e\pi)^{-\frac{1}{4}}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{4\epsilon^*}}.$$

A simple application of Lemma 6.3 gives the asymptotic formula of $u^{iv}(0)$. The asymptotic formula of $u''(0)$ follows from $\epsilon^* u^{iv}(0) = -u'(0)u''(0)$. □

Recall from Corollary 5.4 of section 5 that $u'(0) \sim 1 + \epsilon^*$ as $\epsilon^* \to 0$. This and Lemma 7.1 provide the complete information on $u(t)$, and therefore, on $f(\eta, \epsilon)$ at the turning point.

Since also $u'''(0) \sim -\pi^2\Delta^2$ and $\epsilon^* = \frac{\epsilon}{\Delta}$, we have

$$\pi^2\Delta^2 \sim (2e\pi)^{-\frac{1}{4}}\epsilon^{*-\frac{7}{4}}e^{-\frac{1}{4\epsilon^*}}.$$

It turns out that Δ satisfies

$$\frac{\Delta}{\epsilon}e^{\frac{\Delta}{\epsilon}} \sim \frac{1}{2e\pi^9\epsilon^8}. \tag{7.1}$$

This completes the proof of Theorem 4.5. Thus, from Theorem 3.1,

$$f(\eta) \sim \frac{\sin\frac{\pi\eta}{1-\Delta}}{\frac{\pi\Delta}{1-\Delta}}$$

uniformly on $[0, 1-\Delta]$. The proof of Theorem 2.1 is complete.

Remark 3. The proved asymptotic relation (7.1) between ϵ and Δ is not exactly the same as the formula formally obtained in [7]. Therefore, the formal asymptotic technique used in [7] needs to be modified. This will be given in another paper.

Acknowledgment. The author thanks Professor J. B. McLeod for showing him the details of [9].

REFERENCES

[1] A. S. Berman, *Laminar flow in channels with porous walls*, J. Appl. Phys., 24 (1953), pp. 232–1235.

[2] J. F. Brady, *Flow development in a porous channel and tube.* Phys. Fluids, 27 (1984), pp. 1061–1067.

[3] S. M. Cox, *Two-dimensional flow of a viscous fluid in a channel with porous walls*, J. Fluid Mech., 227 (1991), pp. 1–33.

[4] S. P. Hastings, C. Lu, and A. D. MacGillivray, *A boundary value problem with multiple solutions from the theory of laminar flow*, SIAM J. Math. Anal., 23 (1992), pp. 201–208.

[5] P. A. Lagerstrom, *Matched Asymptotic Expansions,* Springer-Verlag, New York, 1989.

[6] C. Lu and A. D. MacGillivray, *Asymptotic behavior of solutions for a similarity equation for laminar flows in rectangular channels with porous walls*, IMA J. Appl. Math., 49 (1992), pp. 139–162.

[7] A. D. MacGillivray and C. Lu, *Asymptotic solution of a laminar flow in a porous channel with large suction: A nonlinear turning point problem*, Meth. Appl. Anal., 1 (1994), pp. 229–248.

[8] C. Lu, *On existence of multiple solutions of a boundary value problem from pipe flow*, Canad. Quart. Appl. Math., 2 (1994), pp. 361–393.

[9] J. B. McLeod, *Laminar flow in a porous channel*, in Asymptotic Beyond All Orders, #284 NATO ASI Series, H. Segur, S. Tanveer, and H. Levine, eds., Plenum Press, New York, 1991.

[10] M. Morduchow, *On laminar flow through a channel or tube with injection: Applications of method of averages*, Quart. Appl. Math., XIV (1957), pp. 361–368.

[11] I. Proudman and K. Johnson, *Boundary-layer growth near a rear stagnation point*, J. Fluid Mech., 12 (1962), pp. 161–168.

[12] W. A. Robinson, *The existence of multiple solutions for the laminar flow in a uniformly porous channel with suction at both walls,* J. Engrg. Math., 10 (1876), pp. 23–40.

[13] F. M. Skalak and C.-Y. Wang, *On the nonunique solutions of laminar flow through a porous tube or channel*, SIAM J. Appl. Math., 34 (1978), pp. 535–544.

[14] K.-G. Shih, *On the existence of solutions of an equation arising in the theory of laminar flow in a uniformly porous channel*, SIAM J. Appl. Math., 47 (1987), pp. 526–533.

[15] R. M. Terrill, *Laminar flow in a uniformly porous channel with large injection*, Aeronaut Q., 16 (1965), pp. 323–332, 26 (1973), pp. 47–354.

[16] M. B. Zaturska, P. G. Drazin, and W. H. H. Banks, *On the flow of a viscous fluid driven along a channel by suction at porous walls*, Fluid Dyn. Res., 4 (1988), pp. 151–178.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1135–1157, September 1997

FREE BOUNDARY FLUID SYSTEMS IN A SEMIGROUP APPROACH AND OSCILLATORY BEHAVIOR[*]

BEN SCHWEIZER[†]

Abstract. We consider the free boundary problem of a liquid drop with viscosity and surface tension. We study the linearized equations with semigroup methods to get existence results for the nonlinear problem. The spectrum of the generator is computed. Large surface tension creates nonreal eigenvalues, and an exterior force results in a Hopf bifurcation. The methods are used to study wind-generated surface waves.

Key words. Hopf bifurcation, viscous fluid, free boundary

AMS subject classifications. 35Q30, 76D33

PII. S0036141096299892

1. Introduction.

We consider two examples of a finite mass of viscous fluid with a free boundary. In contrast to the case of a fixed domain, the fluid is capable of showing damped oscillations. Due to surface tension the surface area carries potential energy, and oscillations correspond to an exchange of energy between its kinetic and its potential form. The time-dependent problem has parabolic and hyperbolic features, and Beale calls it "mixed in character." The work at hand contributes to the study of this dynamical system.

We assume the system to be close to a stationary solution. The theory is written down for an almost spherical liquid drop but applies also to water in a container with periodic lateral boundaries. We rewrite the equations in semigroup form; the study of the generator reveals some nonstandard properties regarding the choice of function spaces, the spectrum, and the resolvent.

In sections 2–4 we derive an existence theory using the language of semigroup theory and maximal regularity results. In sections 5–7 we study qualitative properties of the spectrum of the generator, such as nonreal or (with an external forcing) imaginary eigenvalues. The two parts interact: the existence theory allows the proof of a Hopf bifurcation in two examples.

The existence theory begins with a proof that the spectrum of the generator consists of eigenvalues and is contained in a sector of the complex plane. Therefore, a natural idea is to apply the semigroup theory for sectorial operators. But the analysis of the resolvent shows that the problem does not fit into this framework: the estimates are valid only on a subspace. On the other hand, due to the kinematic boundary condition, the nonlinearity is always contained in the same subspace. We show that the methods of semigroup theory can be adapted and derive an existence result for the linear problem in section 3 and for the nonlinear problem in section 4. The solutions provide differentiable flows on a Banach manifold. This will be the setting to prove a Hopf bifurcation in section 6.

[*] Received by the editors February 23, 1996; accepted for publication (in revised form) May 2, 1996.

http://www.siam.org/journals/sima/28-5/29989.html

[†] Institut für Angewandte Mathematik, Universität Heidelberg, INF 294, 69120 Heidelberg, Germany (ben.schweizer@iwr.uni-heidelberg.de). This research was done in part while the author visited the Courant Institute for Mathematical Sciences, New York, supported by the Deutsche Forschungsgemeinschaft.

Regarding other works, we wish to mention first Beale [2], who studied infinite domains. He derives the resolvent estimate with the help of Fourier transforms. His proof is considerably longer than ours, since the infinite domain corresponds to a continuous spectrum.

Concerning finite domains, we refer to the numerous works of Solonnikov, who gave the first existence result. He derives estimates with methods from potential theory, after having transformed the equations in a half-space [15]. We mention a related article on Hopf bifurcation in a two-phase fluid system by Renardy and Joseph [13]. In their more physical model the transversal crossing of eigenvalues appears as an assumption. They do not treat the initial value problem. For a more geometric approach, see Bemelmans [4] and Wagner [17].

In the second part of this article we prove qualitative properties of the spectrum. We describe the basic idea for the case where we have only one physical parameter, a nondimensional surface tension β. We are interested in how an eigenvalue λ of the operator $\mathcal{L}_\beta$ depends on β. We do this indirectly. For some function $\tilde{\beta}$ every complex number λ is an eigenvalue of the operator $\mathcal{L}_{\tilde{\beta}(\lambda)}$. The study of the function $\tilde{\beta}$ on the real axis gives us insight into the behavior of $\lambda(\beta)$.

In section 5, we see that for vanishing surface tension the spectrum of $\mathcal{L}$ consists of both the Stokes eigenvalues in a fixed domain and the interfacial eigenvalue 0. For a fixed interfacial eigenmode with increasing β, the first two eigenvalues move towards each other and must leave the real axis, while the other eigenvalues remain trapped in fixed intervals.

We use the same general idea in two examples with an exterior force. In section 6, we assume that the liquid drop experiences negative damping. We can count the eigenvalues inside a ball and prove that if the force reaches a critical strength a pair of eigenvalues crosses transversally the imaginary axis. Using the existence results, we can prove a Hopf bifurcation.

In section 7 we apply the idea to a model for the generation of water waves by wind. A strong wind leads to a Hopf bifurcation. We also gain insight into the shape of the eigenfunctions for strong wind: two of them show the structure of an ideal fluid; the other, approximate Stokes eigenfunctions. There are works (e.g., [10], [11]) that give asymptotic formulas for the eigenvalues for a vanishing exterior force. Our results confirm their qualitative properties and provide additional mathematical insight and proofs.

2. The liquid-drop equations.

We first collect the nonlinear equations describing a liquid drop. Let Ω be the subdomain of $\mathbb{R}^3$ occupied by liquid. The velocity field and the pressure within the liquid drop are denoted by u and p, respectively. The exterior normal vector of $\partial\Omega$ is denoted by n^Ω; tangential vectors are denoted by τ_i^Ω. We use the dimensionless viscosity ν.

The surface tension will become important. The physical quantity is a number $\beta > 0$: if the surface has a mean curvature $H(\eta)$, then the surface tension creates a pressure $2\beta H(\eta)$.

We introduce the strain-tensor

$$(S_u)_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i)$$

and the product $(S_u) : (S_w) := \sum_{i,j} (S_u)_{ij}(S_w)_{ij}$, and we define additionally

$$S_u^n := n \cdot S_u \cdot n.$$

We deal with a free boundary problem. The unknown functions are not only u and p but also the domain Ω. We assume small perturbations of the unit sphere and parametrize the surface of the liquid drop with a function $\eta : S^2 \to \mathbb{R}$. The domain occupied by liquid is

$$\Omega(t) = \{r\xi \in \mathbb{R}^3 | \, \xi \in S^2, 0 \leq r < 1 + \eta(\xi)\}.$$

The velocity field is a function

$$u(t, .) : \Omega(t) \to \mathbb{R}^3.$$

In the interior, the following Navier–Stokes equations hold:

$$\partial_t u + (u \cdot \nabla)u - \nu\Delta u + \nabla p = 0, \tag{1}$$
$$\nabla \cdot u = 0. \tag{2}$$

The boundary conditions are the geometric condition that η always parametrizes the surface, the additional pressure created by surface tension, the condition of vanishing tangential stress, plus initial conditions:

$$\partial_t \eta + (\partial_\varphi \eta)u_\varphi + (\partial_\vartheta \eta)u_\vartheta = u_r, \tag{3}$$
$$p - 2\nu n^\Omega \cdot S_u \cdot n^\Omega = 2\beta \; H(\eta), \tag{4}$$
$$\tau_i^\Omega \cdot S_u \cdot n^\Omega = 0, \tag{5}$$
$$(u, \eta)(t = 0) = (u_0, \eta_0). \tag{6}$$

Equation (3) can be derived by considering a particle at the boundary with position $(r(t), \varphi(t), \vartheta(t))$ in spherical coordinates using $\eta(t, \varphi(t), \vartheta(t)) = r(t)$.

We will return to these nonlinear equations in section 4. We now give the linearization of the problem in $u = 0$, $\eta = 0$. We replace the domain Ω by the unit ball B^3. n^Ω and τ_i^Ω are replaced by the normal and tangential vectors of the unit sphere, $n(\xi) = \xi$ and τ_i. The radial velocity will now be written as $u_n|_\partial = n \cdot u|_\partial$.

The linearization of the mean curvature of $\partial\Omega = \{\xi(1 + \eta(\xi))| \, \xi \in S^2\}$ is denoted by $-\frac{1}{2}\underline{\Delta}\eta$. With the Laplace–Beltrami operator of the sphere Δ_B, there holds

$$\underline{\Delta} = \Delta_B + 2 \cdot id.$$

The linearized equations are

$$\begin{aligned} u : B^3 \to \mathbb{R}^3, \quad p : B^3 \to \mathbb{R}, &\quad \eta : S^2 \to \mathbb{R}, \\ \frac{d}{dt}u - \nu\Delta u + \nabla p &= 0, \\ \nabla \cdot u &= 0, \\ \frac{d}{dt}\eta &= u_n|_\partial, \\ (-p + 2\nu S_u^n)|_\partial &= \beta\underline{\Delta}\eta, \\ (\tau_i \cdot S_u \cdot n)|_\partial &= 0. \end{aligned}$$

In the equation for the pressure, we omitted the constant pressure induced by the surface tension of the unit sphere.

Before we start the analysis of the linearized liquid-drop equations, we collect some facts concerning the Stokes equation. We write $H^r = H^{r,2}$ for Sobolev spaces.

If the domain is the unit ball $B = B^3$, we often omit this argument: $H^r = H^r(B^3)$. The Stokes operator $A : (u, p) \mapsto (-\nu\Delta u + \nabla p, \nabla \cdot u)$ is elliptic in the sense of Agmon, Douglis, and Nirenberg [1] with any of the boundary conditions

$$\begin{aligned} & u|_\partial = 0, \\ \text{or} \quad & u_n|_\partial = 0, \quad \tau \cdot S_u|_\partial \cdot n = 0, \\ \text{or} \quad & (p - 2\nu S_u^n)|_\partial = 0, \quad \tau \cdot S_u|_\partial \cdot n = 0. \end{aligned}$$

Solutions to inhomogeneous boundary data have maximal regularity.

We introduce the operator

$$\mathcal{H} : H^{r-1/2}(S^2) \to H^r(B^3),$$

which maps a function to its harmonic extension.

We will use integration by parts in the following form.

LEMMA 2.1. *For smooth functions $u, w : B^3 \to \mathbb{R}^3$ with $\nabla \cdot u = \nabla \cdot w = 0$ and $\tau \cdot S_u \cdot n = 0$,*

$$2 \int_B S_u : S_w = \int_B \{-\Delta u + \nabla \mathcal{H}(2S_u^n)\} \cdot w.$$

We now return to the liquid-drop equations. Our aim is to write the linear liquid-drop equations in the form $\frac{d}{dt}x + \mathcal{L}x = 0$ and to satisfy the boundary conditions by the choice of appropriate function spaces.

We start by rewriting the boundary condition for the pressure. The pressure p is a harmonic function; therefore

$$p = \mathcal{H}(2\nu S_u^n|_\partial) - \mathcal{H}(\beta \underline{\Delta} \eta).$$

The physical quantities volume, momentum, and angular momentum are conserved. We use this fact in the definition of the function spaces.

The following point of view is useful: the first eigenspace of Δ_B corresponds to constant functions $\Phi(x) = a$; the second eigenspace, to translations $\Phi(x) = b \cdot x$. The physical conditions imply that the projection of η onto the first two eigenspaces of Δ_B vanishes.

DEFINITION 2.2. *Define the Hilbert spaces*

$$\begin{aligned} Y^r &:= \left\{ u \in H^r(B^3)^3 | \nabla \cdot u = 0; \int_{B^3} u = 0; \int_{B^3} u \wedge \gamma = 0 \; \forall \gamma \in \mathbb{R}^3 \right\}, \\ X^r &:= \left\{ (u, \eta) \in Y^r \times H^{r+1-1/2}(S^2) | \int_{S^2} \eta = 0; \int_{S^2} n \cdot \eta = 0 \right\}, \\ \tilde{X}^r &:= \{(u, \eta) \in X^r | n \cdot S_u(z) \cdot \tau|_\partial = 0 \; \forall \tau \in T_z S^2\}, \end{aligned}$$

and the operator

$$\mathcal{L} : X^r \to X^r, \qquad \tilde{X}^r \supset \mathcal{D}(\mathcal{L}) \supset \tilde{X}^{r+2}$$

by

$$\mathcal{L} \begin{pmatrix} u \\ \eta \end{pmatrix} := \begin{pmatrix} -\nu\Delta u + \nabla \mathcal{H}(2\nu S_u^n|_\partial) - \nabla \mathcal{H}(\beta \underline{\Delta} \eta) \\ -u_n|_\partial \end{pmatrix}.$$

Easy calculations show that $\mathcal{L}$ maps to X: the liquid drop does not start to move its center of mass, it does not start to rotate, and it keeps its volume.

The linearized liquid-drop equation reads

$$\frac{d}{dt}x + \mathcal{L}x = 0, \quad x \in \tilde{X}. \tag{7}$$

In the case of a pure rotation the integral $\int_B |S_u|^2$ vanishes without u being a constant. But in our function spaces a Korn inequality holds: there exists a constant C_K such that for all $u \in Y$

$$\frac{1}{C_K}\|Du\|^2_{L^2} \le 2\nu \int_{B^3} |S_u|^2 \le C_K \|Du\|^2_{L^2}. \tag{8}$$

See, e.g., [16]. Here $\|Du\|_{L^2}$ may be replaced by $\|u\|_{H^1}$ because the mean of u vanishes.

DEFINITION 2.3 (energy norms). *For functions $u, v : B^3 \to \mathbb{R}^3$ and $\eta, \sigma : S^2 \to \mathbb{R}$ we define*

$$\begin{aligned} \langle u, v\rangle_E &:= \int_B \bar{u}\cdot v, \\ \langle \eta, \sigma\rangle_E &:= \beta \int_S (-\underline{\Delta}\bar{\eta})\cdot \sigma, \\ \left\langle \begin{pmatrix} u \\ \eta \end{pmatrix}, \begin{pmatrix} v \\ \sigma \end{pmatrix} \right\rangle_E &:= \langle u, v\rangle_E + \langle \eta, \sigma\rangle_E. \end{aligned}$$

The corresponding norms are denoted by $\|.\|_E$.

LEMMA 2.4 (position of eigenvalues of $\mathcal{L}$). *Let $(u, \eta) \in \tilde{X}^2$ be an eigenvector of $\mathcal{L}$ with eigenvalue μ. Then*

$$\begin{aligned} \mathrm{Re}(\mu)\|(u,\eta)\|^2_E &= 2\nu \int_B |S_u|^2, \\ \mathrm{Im}(\mu)\|(u,\eta)\|^2_E &= 2\beta \mathrm{Im}\left(\int_S u_n|_\partial \underline{\Delta}\bar{\eta}\right). \end{aligned}$$

In the case of $\mathrm{Im}(\mu) \neq 0$ the following energy equality holds:

$$\|u\|^2_E = \|\eta\|^2_E = \frac{1}{2}\|(u,\eta)\|^2_E.$$

Proof.

$$\begin{aligned} \mu \left\| \begin{pmatrix} u \\ \eta \end{pmatrix} \right\|^2_E &= \left\langle \begin{pmatrix} u \\ \eta \end{pmatrix}, \mathcal{L}\begin{pmatrix} u \\ \eta \end{pmatrix} \right\rangle_E \\ &= \left\langle \begin{pmatrix} u \\ \eta \end{pmatrix}, \begin{pmatrix} -\nu\Delta u + \nabla\mathcal{H}(2\nu S^n_u) - \nabla\mathcal{H}(\beta\underline{\Delta}\eta) \\ -u_n|_\partial \end{pmatrix} \right\rangle_E \\ &= \int_B \{\bar{u}\cdot(-\nu\Delta u + \nabla\mathcal{H}(2\nu S^n_u))\} \\ &\quad - \int_B \bar{u}\nabla\mathcal{H}(\beta\underline{\Delta}\eta) - \beta\int_S (-\underline{\Delta}\bar{\eta})(u_n)|_\partial \\ &= 2\nu\int_B |S_u|^2 + \beta\int_S \{u_n|_\partial \underline{\Delta}\bar{\eta} - \bar{u}_n|_\partial \underline{\Delta}\eta\}. \end{aligned}$$

This implies the assertion on the real and the imaginary part of μ.

To prove the energy equality we use the second part of the eigenvalue equation, $-u_n|_\partial = \mu\eta$:

$$\begin{aligned}\mathrm{Im}(\mu)\left\|\begin{pmatrix} u \\ \eta \end{pmatrix}\right\|_E^2 &= 2\beta \mathrm{Im}\left(\int_S u_n|_\partial \underline{\Delta}\bar{\eta}\right)\\ &= 2\mathrm{Im}(\mu)\|\eta\|_E^2. \qquad \square\end{aligned}$$

Using the properties of the Stokes operator one easily proves the following lemma for $\beta \neq 0$.

Lemma 2.5. *The operator $\mathcal{L}^{-1} : X^r \to \tilde{X}^{r+1}$ is bounded.*

We point out that $\mathcal{L}^{-1} : X^0 \to \tilde{X}^2$ is not bounded: let (u, η) solve $\mathcal{L}(u, \eta) = (0, g)$. A bound for $\|u\|_{H^2}$ would imply $g = u_n|_\partial \in H^{3/2,2}(S^2)$. But a priori, only $g \in H^{1/2,2}(S^2)$ holds.

The Lumer–Phillips theorem implies the following lemma.

Lemma 2.6. *$\mathcal{L}$ generates a C^0-semigroup in the space X_E corresponding to the energy norms.*

We want to split X into a direct sum of $\mathcal{L}$-invariant subspaces $(X_n)_{n\in\mathbb{N}}$ according to spherical harmonics. The functions $\{\psi_{n,k}|n \in \mathbb{N}, k \in \{-n, ..., n\}\}$ shall span Ψ_n, the nth eigenspace of the Laplace–Beltrami operator of S^2. We denote the corresponding eigenvalue by $\Lambda_n > 0$ and the eigenvalue of $-\underline{\Delta}$ by $\underline{\Lambda_n} = \Lambda_n - 2$. Let ν be the normal vector, ∇_T the tangential gradient, and $\nabla_T^\perp = \nu \wedge \nabla$ the orthogonal tangential gradient.

A vector field is in X_n if on any sphere of radius r the function can be represented by $\psi_{n,k}\, \nu$, $\nabla_T \psi_{n,k}$, and $\nabla_T^\perp \psi_{n,k}$, $k \in \{-n, ..., n\}$.

Proposition 2.7. *The spectrum of $\mathcal{L}$ consists only of eigenvalues which are contained in a sector*

$$S_C := \{\mu| \quad |\mathrm{Im}(\mu)| < C\mathrm{Re}(\mu)\}. \tag{9}$$

Proof. By Lemma 2.5, $\mathcal{L}$ has a compact resolvent and therefore a pure point spectrum. We prove that eigenvalues are contained in a sector S_C. Let μ be an eigenvalue with eigenvector $(u, \eta) \in \tilde{X}_k$. If $\mathrm{Im}(\mu) = 0$, then μ is contained in any sector S_C. We can therefore assume $\mathrm{Im}(\mu) \neq 0$.

We use Lemma 2.4 and the fact that u depends on the radius r like $e^{\sqrt{\underline{\Lambda_k}+|\mu|}\ r}$.

$$\begin{aligned}|\mathrm{Im}(\mu)| \ \left\|\begin{pmatrix} u \\ \eta \end{pmatrix}\right\|_E^2 &= 2\beta\left|\mathrm{Im}\left(\int_S u_n|_\partial \underline{\Delta}\bar{\eta}\right)\right|\\ &\leq \beta\int_S |\underline{\Delta}\eta|^2 + \beta\int_S |u_n|_\partial|^2\\ &\leq |\underline{\Lambda_k}|\|\eta\|_E^2 + \beta C_T\sqrt{\underline{\Lambda_k} + |\mu|}\ \|u\|_{L^2(B)}^2.\end{aligned}$$

Therefore,

$$|\mathrm{Im}(\mu)| \leq C_1|\underline{\Lambda_k}| + \beta C_2 \mathrm{Re}(\mu).$$

Using Lemma 2.4 and the Korn inequality yields

$$\mathrm{Re}(\mu)\left\|\begin{pmatrix} u \\ \eta \end{pmatrix}\right\|_E^2 = 2\nu\int_B |S_u|^2$$

$$\geq \frac{1}{C_K}\|\nabla_T u\|_{L^2}^2$$
$$\geq \frac{1}{2C_K}|\underline{\Lambda_k}|\left\|\begin{pmatrix} u \\ \eta \end{pmatrix}\right\|_E^2 .$$

The assertion follows with the constant $C = 2C_1C_K + \beta C_2$. □

3. An estimate for resolvents and the time-dependent problem. In this section we collect estimates for the resolvent. A first type of estimate concerns solutions (u,η) of $(\lambda - \mathcal{L})(u,\eta) = (f,0)$. Such estimates are known in similar contexts [2], [13]. We indicate how they can be derived more easily in our case of only one fluid in a compact domain.

As a corollary, we get a second type of estimate concerning the resolvent on the full space: we show that $\mathcal{L}$ is a sectorial operator.

To solve nonlinear equations, it will be necessary to increase the regularity of the function spaces. In this section we use $X = X^r$ and $\tilde{X}^{++} = \tilde{X}^{2+r}$ with $r = 0$ or $r = 2$.

Theorem 3.1. *There exists $C_R > 0$ such that solutions $(u,\eta) \in \tilde{X}^{++}$ of*

$$(\lambda - \mathcal{L})\begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$$

with $\lambda \in \mathbb{C} \setminus S_C$ satisfy the regularity

$$\|(u,\eta)\|_{X^{++}} \leq C_R\|(f,0)\|_X \tag{10}$$

and the resolvents estimate

$$\|(u,\eta)\|_X \leq C_R\frac{1}{|\lambda|}\|(f,0)\|_X . \tag{11}$$

Proof. We indicate the ideas of the proof. One writes the equation as

$$\lambda u + \nu\Delta u - \nabla\mathcal{H}(2\nu S_u^n) - \frac{1}{\lambda}\nabla\mathcal{H}(\beta\underline{\Delta}u_n|_\partial) = f. \tag{12}$$

Testing with $\Delta_B\bar{u}$ and taking real and imaginary parts shows that $\beta\frac{1}{|\lambda|}\int_{S^2}|\underline{\Delta}u_n|_\partial|^2$ and $\|u\|_{H^2}^2$ can be estimated by $\|f\|_{L^2}$ and $|\lambda|^2\|u\|_{L^2}^2$.

Testing (12) with $\bar{u}$ and taking the imaginary part yields

$$|\lambda|^2\|u\|_{L^2}^2 \leq \beta\int_{S^2}\bar{u}_n\underline{\Delta}u_n + \mathrm{const}\|f\|_{L^2}|\lambda|\|u\|_{L^2}$$
$$\leq \beta C_c\|u\|_{L^2}^{1/2}\|u\|_{H^2}^{3/2} + \mathrm{const}\|f\|_{L^2}^2 + \frac{1}{2}|\lambda|^2\|u\|_{L^2}^2 .$$

This yields the estimates for u. Equation (12) provides the estimates for η. □

Corollary 3.2. *$\mathcal{L}$ is a sectorial operator on $X^{r+2}, r \geq 0$. The spectrum $\sigma(\mathcal{L})$ is contained in a sector S_C, and with a constant $M > 0$, for every $\lambda \in \mathbb{C} \setminus S_C$,*

$$(\lambda - \mathcal{L})\begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \quad\Rightarrow\quad \left\|\begin{pmatrix} u \\ \eta \end{pmatrix}\right\|_{X^{r+2}} \leq \frac{M}{|\lambda|}\left\|\begin{pmatrix} f \\ g \end{pmatrix}\right\|_{X^{r+2}} .$$

Proof. Instead of (12),

$$\lambda u + \nu \Delta u - \nabla \mathcal{H}(2\nu S_u^n) - \frac{1}{\lambda} \nabla \mathcal{H}(\beta \underline{\Delta} u_n|_\partial) = f - \frac{1}{\lambda} \nabla \mathcal{H}(\beta \underline{\Delta} g|_\partial). \tag{13}$$

For $f = 0$ Theorem 3.1 yields

$$\begin{aligned} \|(u,\eta)\|_{X^{r+2}} + |\lambda| \, \|(u,\eta)\|_{X^r} &\leq \text{const} \frac{1}{|\lambda|} \|g\|_{H^{r+3-1/2}(S^2)} \\ &\leq \text{const} \frac{1}{|\lambda|} \|(0,g)\|_{X^{r+2}}. \qquad \square \end{aligned}$$

The preceding corollary verifies one of the assumptions in the Hopf bifurcation theorem of [5]; nevertheless, that theorem cannot be applied since it assumes the nonlinearity to be of lower order.

Using Theorem 3.1 we can solve the initial value problem with the ideas of semigroup theory. Unlike in other approaches, this will provide a time-t map for the nonlinear evolution system. In particular, this tool allows an elementary proof of a Hopf bifurcation theorem.

Definition 3.3. *Let $I := [0,T]$ be a fixed time interval, $T > 0$. We introduce the spaces*

$$Z := C^\alpha(I, X),$$

$$\tilde{Z}^{++} := C^{1,\alpha}(I, X) \cap C^\alpha(I, \tilde{X}^{++}).$$

For the regularity of the initial values we define

$$D_{\mathcal{L},\alpha} := \{x \in X | \quad \|x\|_{K,\alpha} := \|t \mapsto e^{-t\mathcal{L}} x\|_{Z^{++}} < \infty\}.$$

As in the semigroup theory, we choose a path of integration Γ in the complex plane containing S_C and write functions in $\mathcal{L}$ as integrals over Γ. Following the standard lines we get the special semigroup estimates

$$\|e^{-t\mathcal{L}}(f,0)\|_X + \|t\mathcal{L}e^{-t\mathcal{L}}(f,0)\|_X + \|(t\mathcal{L})^2 e^{-t\mathcal{L}}(f,0)\|_X \leq C\|(f,0)\|_X,$$
$$\|\mathcal{L}^{-1}e^{-t\mathcal{L}}(f,0)\|_{X^{++}} + \|te^{-t\mathcal{L}}(f,0)\|_{X^{++}} + \|t^2\mathcal{L}e^{-t\mathcal{L}}(f,0)\|_{X^{++}} \leq C\|(f,0)\|_X.$$

We can now prove a result of maximal regularity, i.e., solutions of $\partial_t x + \mathcal{L}x = F$ are in Z^{++} if F is in Z. The underlying idea is taken from [6], which proves a regularity result in the case when the resolvent of the generator has optimal regularity properties.

Theorem 3.4. *Let $F \in Z = C^\alpha(I, X)$ be of the form $F = \begin{pmatrix} f \\ 0 \end{pmatrix}$ and let $x_0 \in \tilde{X}^{++}$ satisfy the compatibility condition*

$$x_0 - \mathcal{L}^{-1}F(0) \in D_{\mathcal{L},\alpha}.$$

Then the equation

$$\partial_t x + \mathcal{L}x = \begin{pmatrix} f \\ 0 \end{pmatrix},$$
$$x(0) = x_0$$

has a unique solution $x \in \tilde{Z}^{++}$ *bounded by*

$$\|x\|_{Z^{++}} \leq C_1(T, M, \alpha)\{\|x_0\|_{X^{++}} + \|x_0 - \mathcal{L}^{-1}F(0)\|_{K,\alpha} + \|F\|_Z\}.$$

The compatibility condition is always satisfied:

$$\|x(t) - \mathcal{L}^{-1}F(t)\|_{K,\alpha} \leq C_2(T, M, \alpha) \cdot \{\|x_0\|_{X^{++}} + \|x_0 - \mathcal{L}^{-1}F(0)\|_{K,\alpha} + \|F\|_Z\}.$$

Proof. One proves that $x(.)$ is in $C^\alpha(I, X^{++})$ by decomposing $x(t)$ as

$$\begin{aligned} x(t) &= e^{-t\mathcal{L}}x_0 + \int_0^t e^{-(t-s)\mathcal{L}}F(s)ds \\ &= e^{-t\mathcal{L}}(x_0 - \mathcal{L}^{-1}F(0)) + e^{-t\mathcal{L}}\mathcal{L}^{-1}(F(0) - F(t)) \\ &\quad + \int_0^t e^{-(t-s)\mathcal{L}}(F(s) - F(t))ds + \mathcal{L}^{-1}F(t). \end{aligned}$$

The first term is in $C^\alpha(I, X^{++})$ by the compatibility condition; for the other terms we can use the special semigroup estimates. The compatibility condition in t is proved in a similar manner. □

We complete our analysis of the nonstationary equation with a remark about the size of the space $D_{\mathcal{L},\alpha}$. The preceding theorem shows that it contains all functions $(u(t), \eta(t))$ that can be reached with solutions starting from 0. It furthermore contains all smooth functions with appropriate boundary data. By writing $e^{-t\mathcal{L}}x_0 - x_0$ as a complex integral one proves, for $X = X^r$,

$$\{x_0 \in \tilde{X} | \mathcal{L}x_0 \in \tilde{X}^{r+4}\} \subset D_{\mathcal{L},\alpha}.$$

Concerning the time-dependent problem, we finally remark that the equations for the center of mass and rotations can now be solved with one additional integration.

4. The nonlinear liquid-drop equation. We now consider the full free boundary problem. We will need high orders of regularity and set $X = X^2, X^{++} = X^4$. In this section we do not impose the conditions of vanishing momentum and vanishing angular momentum and extend $\mathcal{L}^{-1}$ trivially.

The transformation of the equations (1)–(6) to a fixed domain is done in the standard way, as in Beale [2]. Using the operator $\mathcal{L}$, the transformed equations read

(14) $$\partial_t \begin{pmatrix} v \\ \eta \end{pmatrix} + \mathcal{L}\begin{pmatrix} v \\ \eta \end{pmatrix} = \begin{pmatrix} F(v,\eta) \\ 0 \end{pmatrix}$$

with the boundary condition

(15) $$\tau_i \cdot S_v \cdot n = G_i(v, \eta).$$

In the case $r \geq 1$,

$$F : X^{r+2} \to H^r(B^3)^3,\ F(0,0) = 0,\ DF \text{ exists, and } DF(0,0) = 0,$$

$$G : X^{r+2} \to H^{r+1-\frac{1}{2}}(S^2)^2,\ G(0,0) = 0,\ DG \text{ exists, and } DG(0,0) = 0.$$

We solve the time-dependent problem by means of an iteration. The boundary condition (15) is satisfied with the help of a function Φ.

Definition 4.1. *For functions g_i we define a vector field $\Phi(g) : B^3 \to \mathbb{R}^3$ which has the correct boundary values. With the help of the Stokes operator A we define $\Phi(g) : B^3 \to \mathbb{R}^3$ as the solution of*

$$\begin{aligned} A\Phi(g) &= 0 \textit{ in } B^3,\\ \Phi_n(g) &= 0 \textit{ on } S^2,\\ \tau_i \cdot S_{\Phi(g)} \cdot n &= g_i \textit{ on } S^2. \end{aligned}$$

We consider new variables, namely,

$$\tilde{x} = x - \begin{pmatrix} \Phi(G(x)) \\ 0 \end{pmatrix} \tag{16}$$

with inverse $x = \xi(\tilde{x})$. The boundary condition (15) is satisfied if we construct $\tilde{x} \in \tilde{X}$. In the $\tilde{x}$-variable the equations read

$$(\partial_t + \mathcal{L})\tilde{x} = \begin{pmatrix} \tilde{F}(\tilde{x}) \\ 0 \end{pmatrix} := \begin{pmatrix} F \circ \xi(\tilde{x}) \\ 0 \end{pmatrix} - (\partial_t + \mathcal{L}) \begin{pmatrix} \Phi \circ G \circ \xi(\tilde{x}) \\ 0 \end{pmatrix},$$

$$\tilde{x}(0) = \xi^{-1}(x_0) = x_0 - (\Phi \circ G(x_0), 0). \tag{17}$$

We remark that we have a vanishing second component in the right-hand side.

In the following, we impose the conditions that initial values satisfy (5) and that the formal time derivative at 0 has the appropriate regularity.

Definition 4.2. *x_0 satisfies the nonlinear compatibility conditions in X^{++} if*

$$\tilde{x}_0 := x_0 - \begin{pmatrix} \Phi(G(x_0)) \\ 0 \end{pmatrix} \in \tilde{X}^{++}, \tag{18}$$

$$z := \tilde{x}_0 - \mathcal{L}^{-1}(\tilde{F}(\tilde{x}_0), 0) \in D_{\mathcal{L},\alpha}. \tag{19}$$

Before we state the theorem of local existence and uniqueness, we investigate the compatibility conditions in more detail. The following proposition states that the permitted small initial values form a Banach manifold. We will use this fact to prove a Hopf bifurcation; the idea is taken from Koch [9].

Proposition 4.3. *There exists $U = B_\varepsilon(0) \subset D_{\mathcal{L},\alpha}$ and a mapping $\zeta : U \to X^{++}$ such that every $x_0 = \zeta(z)$ satisfies the nonlinear compatibility conditions with small norms. We denote the manifold $\zeta(U)$ by $\mathcal{M}$. ζ can be constructed with $D\zeta(0) = id$.*

Proof. We only have to invert the equation (19) with $x_0 \in \tilde{X}^{++}$. We use the contraction mapping principle for the map

$$\tilde{X}^{++} \ni \tilde{x}_0 \mapsto z + \mathcal{L}^{-1}(\tilde{F}(\tilde{x}_0), 0) \in \tilde{X}^{++}.$$

This yields the fixed point $\tilde{x}_0 = \tilde{\zeta}(z)$. We define $\zeta := \xi \circ \tilde{\zeta}$. □

Theorem 4.4. *According to small initial values x_0 satisfying the compatibility conditions, i.e.,*

$$x_0 = \zeta(z) \in \mathcal{M}, \quad \|z\|_{K,\alpha} \textit{ small},$$

there exists a unique small solution of the nonlinear liquid-drop equation in

$$Z^{++} = C^{1,\alpha}(I, X^2) \cap C^\alpha(I, X^4).$$

As a map on the Banach manifold $\mathcal{M}$, the flow is differentiable.

Proof. We construct the solution with the help of an iteration map $T : \tilde{Z}^{++} \to \tilde{Z}^{++}$. Let $(v, \sigma) \in \tilde{Z}^{++}$ be given. We define $\tilde{f} = \tilde{F}(v, \sigma)$ and solve

$$(\partial_t + \mathcal{L})\tilde{x} = (\tilde{f}, 0),$$

$$\tilde{x}(0) = x_0 - (\Phi \circ G(x_0), 0) = \tilde{\zeta}(z)$$

for $\tilde{x} \in \tilde{Z}^{++}$ with the help of Theorem 3.4. Because of $DF(0,0) = 0$ and $DG(0,0) = 0$, the solution operator is contracting in a small ball $B_\varepsilon(0) \subset Z^{++}$ and there exists a unique fixed point $\tilde{x}$. The function $\xi(\tilde{x})$ is a solution of the nonlinear equation.

We have to take care that in the iteration the right-hand side is contained in the function space. The condition of vanishing divergence can be assured with the usual projection. This yields an additional pressure that vanishes at the boundary.

It remains to show the differentiability of the flow $x_0 \mapsto x(t)$ on the manifold, i.e., the differentiability of

$$\begin{aligned}\Phi_t : \bar{x}_0 \mapsto & \zeta^{-1} x(t) = x(t) - \mathcal{L}^{-1}\tilde{F}(x(t)) \quad \text{with} \\ & \bar{x}_0 = \zeta^{-1}(x_0) = x_0 - \mathcal{L}^{-1}\tilde{F}(x_0) = \bar{x}_0.\end{aligned}$$

We omit the straightforward calculation, proving that the derivative of Φ_t can be written as

$$D\Phi_t : \bar{w}_0 \mapsto (id - \mathcal{L}^{-1} \circ D\tilde{F}|_{x(t)})w(t),$$

where $w(.)$ solves

$$(\partial_t + \mathcal{L})w(.) = D\tilde{F}|_{x(.)} \cdot w(.),$$

$$w(0) = w_0 = (id - \mathcal{L}^{-1} \circ D\tilde{F}|_{x(t)})^{-1}\bar{w}_0. \qquad \square$$

5. The spectrum of $\mathcal{L}$. The eigenvalues of $\mathcal{L}$ can be calculated explicitly for $\beta = 0$. We investigate the movement of the eigenvalues in the complex plane as $\beta \to \infty$. We prove the qualitative behavior that has been observed numerically (compare [3]).

We will make fundamental use of the fact that the liquid-drop problem has an $O(3)$ symmetry, in other words, that $\mathcal{L}$ is $O(3)$-equivariant. The group action will be denoted by $*$. We use the decomposition $X = \bigoplus X_n$, and we describe the spectrum on X_n for $n \geq 2$.

The nth eigenspace of Δ_B, Ψ_n has the standard basis $(\psi_{n,-n}, ..., \psi_{n,n})$ with

$$\psi_{n,k}(\theta, \varphi) = P_{n,k}(cos(\theta))e^{ik\varphi}.$$

The function $\Phi_0 := \psi_{n,0}$ has an isotropy subgroup Γ isomorphic to $O(2)$. Any function $\psi : S^2 \to \mathbb{R}$ can be Γ-symmetrized by

$$\bar{\psi}(\xi) := \fint_\Gamma \gamma * \psi(\xi) d\gamma.$$

The same can be done with functions $v : B^3 \to \mathbb{R}^3$.

In this and the following section we consider only eigenfunctions $(u, \eta) \in X_n$ with $\eta = \Phi_0$. This is no restriction since every eigenfunction can be projected and symmetrized such that the second component is a multiple of Φ_0.

We make constant use of the following observation: given an eigenvalue z of $\mathcal{L}$, we can construct the eigenfunction (u, Φ_0) as the solution of a Stokes problem.

Definition 5.1. *By A_N we denote the Stokes operator in the space of functions with vanishing normal component at the boundary. The eigenvalues of A_N are denoted by $\{\kappa_j\}_{j\in\mathbb{N}}$. The corresponding eigenfunctions with symmetry Γ are denoted by $\{u_j\}_{j\in\mathbb{N}}$; and the pressure, by $\{p_j\}_{j\in\mathbb{N}}$. Their signs are determined in* (24).

For $z \in \mathbb{C} \setminus \{\kappa_j | j \in \mathbb{N}\}$ we define $(\tilde{u}(z), \tilde{p}(z))$ as the unique solution of the system

$$\begin{aligned} z\tilde{u}(z) + \nu\Delta\tilde{u}(z) - \nabla\tilde{p}(z) &= 0,\\ \nabla\cdot\tilde{u}(z) &= 0,\\ \tau\cdot S_{\tilde{u}(z)}|_\partial\cdot n &= 0,\\ \tilde{u}_n(z)|_\partial &= -z\Phi_0. \end{aligned} \tag{20}$$

The solution has the same symmetry as Φ_0, i.e., Γ. In particular, $(\tilde{p}(z) - 2\nu S^n_{\tilde{u}(z)})|_\partial$ has the symmetry Γ and is a multiple of Φ_0. We define $\tilde{r}(z) \in \mathbb{C}$ by

$$(\tilde{p}(z) - 2\nu n\cdot S_{\tilde{u}(z)}\cdot n)|_\partial =: \tilde{r}(z)\Phi_0. \tag{21}$$

We remark that $z \in \mathbb{R}$ implies $\tilde{r}(z) \in \mathbb{R}$.

Any $z \in \mathbb{C} \setminus \{\kappa_j | j \in \mathbb{N}\}$ is an eigenvalue of $\mathcal{L}$ with eigenfunction $(\tilde{u}(z), \Phi_0)$ if it satisfies

$$\tilde{r}(z) = \underline{\Lambda_k}\beta. \tag{22}$$

We remark that the functions $\tilde{u}(z)$ and $\tilde{p}(z)$ can be computed explicitly in terms of Bessel functions; this can be used to analyze the function $\tilde{r}(z)$ numerically. In the following, $\|.\|$ denotes the L^2-norm.

Proposition 5.2 (properties of $\tilde{u}(z)$). *$\tilde{u}(z)$ is a differentiable family of functions for $z \in \mathbb{C} \setminus \{\kappa_j | j \in \mathbb{N}\}$. In κ_j,*

$$\|\tilde{u}(z)\| \to \infty \quad \textit{for} \quad z \to \kappa_j. \tag{23}$$

The rescaled functions approximate the Stokes eigenfunctions

$$u_j = \lim_{\mathbb{R}\ni z\nearrow\kappa_j} \frac{\tilde{u}(z)}{\|\tilde{u}(z)\|} = \lim_{\mathbb{R}\ni z\searrow\kappa_j} \frac{-\tilde{u}(z)}{\|\tilde{u}(z)\|}. \tag{24}$$

Furthermore,

$$\|\tilde{u}(z)\| \to \infty \quad \textit{for} \quad |z| \to \infty. \tag{25}$$

Proof. We define a family of functions $(u(z), p(z))$ which depends smoothly on z in a neighborhood of κ_j by solving

$$\begin{aligned} zu(z) + \nu\Delta u(z) - \nabla p(z) &= 0,\\ \nabla\cdot u &= 0,\\ \tau\cdot S_{u(z)}\cdot n|_\partial = 0, \quad (p(z) - 2\nu S^n_{u(z)})|_\partial &= \Phi_0. \end{aligned}$$

The solution is unique; therefore, $u(\kappa_j)$ is a multiple of u_j. With the notation

$$u_n(z)|_\partial =: s(z)\Phi_0,$$

$s(.)$ is continuous and $s(\kappa_j) = 0$. (23) is proved by observing

$$\tilde{u}(z) = \frac{-z}{s(z)} u(z).$$

(24) is proved by showing that the function $s(.)|_{\mathbb{R}}$ changes sign in κ_j.

Assume $\partial_z s(\kappa_j) = 0$. Then $v := \partial_z u(\kappa_j), q := \partial_z p(\kappa_j)$ satisfies the boundary conditions

$$\tau \cdot S_v|_\partial \cdot n = 0, \quad v_n|_\partial = 0, \quad (q - 2\nu S_v^n)|_\partial = 0.$$

Additionally,

$$\kappa_j v + \nu \Delta v - \nabla q = -u(\kappa_j).$$

Multiplying with $u(\kappa_j)$ and integrating yields $0 = -\|u(\kappa_j)\|^2$—a contradiction.

(25) can be proved directly. As the solution of the Stokes system (20), $\tilde{u}$ satisfies an estimate

$$\|\tilde{u}(z)\|_{H^2} \leq C_S \, \{|z| \, \|\tilde{u}(z)\|_{L^2} + |z| \, \|\Phi_0\|_{H^{3/2}(S^2)}\}.$$

We now use $\tilde{u}_n(z)|_\partial = -z\Phi_0$, a trace formula, and an interpolation to calculate

$$\begin{aligned} |z|^2 \|\Phi_0\|^2_{L^2(S^2)} &= \|\tilde{u}_n(z)|_\partial\|^2_{L^2(S^2)} \\ &\leq C_T \|\tilde{u}(z)\|^2_{H^1} \\ &\leq C_T C_c \|\tilde{u}(z)\|_{L^2} \|\tilde{u}(z)\|_{H^2} \\ &\leq C_T C_c C_S \|\tilde{u}(z)\|_{L^2} \{|z| \|\tilde{u}(z)\|_{L^2} + |z| \|\Phi_0\|_{H^{3/2}(S^2)}\}. \end{aligned}$$

This yields $\|\tilde{u}(z)\|^2_{L^2} \geq \text{const} \cdot |z|$, and the proposition is proved. □

PROPOSITION 5.3 (properties of $\tilde{r}(z)$). *The function $\tilde{r}(z)$ satisfies*

$$\begin{aligned} &\tilde{r}(z) \to 0 \text{ for } \mathbb{R} \ni z \searrow 0, \\ &\tilde{r}(z) \to -\infty \text{ for } \mathbb{R} \ni z \nearrow \kappa_j, \\ &\tilde{r}(z) \to +\infty \text{ for } \mathbb{R} \ni z \searrow \kappa_j. \end{aligned} \tag{26}$$

$\tilde{r}(z)$ is positive for small $z > 0$,

$$\partial_z \tilde{r}(0) > 0. \tag{27}$$

Between κ_j and κ_{j+1}, there is at most one turning point. Critical values of $\tilde{r}(z)$ are positive.

Proof. The assertion of (26) for $z \to 0$ is trivial. u_j satisfies

$$u_j = \lim_{z \nearrow \kappa_j} \frac{\tilde{u}(z)}{\|\tilde{u}(z)\|}.$$

We test the eigenvalue equation of u_j with $v := \frac{\tilde{u}(\kappa_j - \varepsilon)}{\|\tilde{u}(\kappa_j - \varepsilon)\|}$ to get

$$\begin{aligned} 0 &= \langle (\kappa_j - A) u_j, v \rangle \\ &= \langle u_j, (\kappa_j - A) v \rangle - \int_S (p_j - 2\nu S^n_{u_j})|_\partial v_n|_\partial \\ &= \varepsilon \, \langle u_j, v \rangle - \int_S (p_j - 2\nu S^n_{u_j})|_\partial v_n|_\partial. \end{aligned}$$

By (24) the first term is positive for small $|\varepsilon|$; the second has the sign of $\tilde{r}(\kappa_j - |\varepsilon|)$.

To prove (27) we consider the functions $v := \partial_z \tilde{u}(0)$, $q = \partial_z \tilde{p}(0)$. They solve

$$\nu \Delta v - \nabla q = 0,$$

$$\tau \cdot S_v|_\partial \cdot n = 0, \quad v_n|_\partial = -\Phi_0, \quad (q - 2\nu S_v^n)|_\partial = \partial_z \tilde{r}(0)\Phi_0.$$

Multiplying this equation with v yields

$$-\int_{B^3} |S_v|^2 + \partial_z \tilde{r}(0)\|\Phi_0\|^2 = 0,$$

which proves (27).

We claim that turning points of $\tilde{r}(z), z \in \mathbb{R}$, are the critical points of $\|\tilde{u}(z)\|$, $z \in \mathbb{R}$. We consider the functions $v(z) := \partial_z \tilde{u}(z)$, $q(z) := \partial_z \tilde{p}(z)$, which solve

$$\begin{aligned} &\tilde{u}(z) + z v(z) + \nu \Delta v(z) - \nabla q(z) = 0, \\ &\tau \cdot S_{v(z)}|_\partial \cdot n = 0, \quad v_n(z)|_\partial = -\Phi_0, \\ &(q(z) - 2\nu S^n_{v(z)})|_\partial = \partial_z \tilde{r}(z)\Phi_0. \end{aligned} \tag{28}$$

Multiplying (28) by $\tilde{u}(z)$ yields

$$\|\tilde{u}(z)\|^2 + z\partial_z \tilde{r}(z)\|\Phi_0\|^2 - \tilde{r}(z)\|\Phi_0\|^2 = 0, \tag{29}$$

and differentiating gives

$$\partial_z \|\tilde{u}(z)\|^2 + z\partial_z^2 \tilde{r}(z)\|\Phi_0\|^2 = 0. \tag{30}$$

Consider $w := \partial_z^2 \tilde{u}(z_0)$, $r := \partial_z^2 \tilde{p}(z_0)$, which solve

$$\begin{aligned} &2v + \kappa_j w + \nu \Delta w - \nabla r = 0, \\ &\tau \cdot S_w|_\partial \cdot n = 0, \quad w_n|_\partial = 0, \\ &(r - 2\nu S^n_w)|_\partial = \partial_z^2 \tilde{r}(z_0)\Phi_0. \end{aligned} \tag{31}$$

Multiplying (31) with v yields

$$2\|v\|^2 - \langle \tilde{u}(z_0), w\rangle + \partial_z^2 \tilde{r}(z_0)\|\Phi_0\|^2 = 0. \tag{32}$$

Assume there is more than one point with $\partial_z^2 \tilde{r} = 0$. By (30) they coincide with critical points of $\|\tilde{u}(z)\|$. At least one of them satisfies $\partial_z^2 \|\tilde{u}(z)\|^2 \leq 0$. This contradicts equation (32), which implies

$$\partial_z^2 \|\tilde{u}(z)\|^2 = 2\,\langle \tilde{u}(z_0), w\rangle + 2\|v\|^2 = 6\|v\|^2 > 0.$$

The last assertion of the proposition follows from (29). □

We denote the Stokes operator with boundary condition $(p - 2\nu S^n_u)|_\partial = 0$ by A_S and the eigenvalues of A_S by $\{\rho_j\}_{j\in\mathbb{N}}$.

THEOREM 5.4 (the spectrum of $\mathcal{L}$ in dependence of β). *It holds that*

$$\rho_0 < \kappa_0 < \rho_1 < \cdots < \rho_j < \kappa_j < \cdots .$$

For $\beta = 0$ all the eigenvalues of $\mathcal{L}_\beta$ are real. Denoting them by $(\mu_j)_{j\in\mathbb{N}}$,

$$\mu_0 = 0, \quad \mu_{j+1} = \rho_j.$$

For small β the eigenvalues stay real. With increasing β the first eigenvalue moves to the right while the other eigenvalues move to the left. For some $\beta_0 > 0$ the first two eigenvalues merge and leave the real axis.

Given a radius k there exists $\beta_k > 0$ such that for $\beta > \beta_k$ the following is true. The norm of nonreal eigenvalues of $\mathcal{L}_\beta$ is larger than k. Every interval $[\kappa_j, \kappa_{j+1}]$ with $\kappa_{j+1} < k$ contains one and only one eigenvalue $\mu(\beta)$ of $\mathcal{L}_\beta$. This eigenvalue satisfies

$$\mu(\beta) \searrow \kappa_j \textit{ for } \beta \to \infty.$$

Proof. The numbers $\rho_j, j \in \mathbb{N}$ are the zeros of $\tilde{r}(z)$. The shape of $\tilde{r}$ implies the assertion on the position of the Stokes eigenvalues. For $\beta = 0$ we can compute a complete set of eigenfunctions in X_k: $\mu_0 = 0$ with eigenfunction $(0, \Phi_0)$ and $\mu_{j+1} = \rho_j$ with eigenfunctions $(\tilde{u}(\rho_j), \Phi_0)$. By the shape of $\tilde{r}(z)$ and (22) the first two eigenvalues meet at the maximum of $\tilde{r}$ and must leave the real axis. Eigenfunctions $(u(\beta), \Phi_0)$ of $\mathcal{L}_\beta$ with nonreal eigenvalues $\mu(\beta)$ satisfy the energy equality

$$\|u\|_{L^2} = \beta \underline{\Lambda_k} \|\Phi_0\|^2_{L^2(S^2)}.$$

Therefore, nonreal eigenvalues cannot stay bounded for $\beta \to \infty$. The shape of $\tilde{r}$ together with (22) prescribes the movement of the real eigenvalues as stated. The theorem is proved. □

REMARK 5.5. *Eigenvalues leave the real axis with an infinite speed. The qualitative shape of $\tilde{r}_\nu(z)$ is independent of the viscosity ν:*

$$\tilde{r}_{\alpha\nu}(\alpha z) = \alpha^2 \tilde{r}_\nu(z).$$

Proof. Eigenvalues leave the real axis in a critical point z_0 of $\tilde{r}$, an analytic function in $\mathbb{C} \setminus \{\kappa_j | j \in \mathbb{N}\}$. It holds that $\frac{\partial \mathrm{Re}(\tilde{r}(z_0))}{\partial \mathrm{Im} z} = 0$, and (22) implies

$$\underline{\Lambda_k} = \frac{\partial \mathrm{Re}(\tilde{r}(z(\beta)))}{\partial \beta} = \frac{\partial \mathrm{Re}(\tilde{r}(z))}{\partial z} \frac{\partial z}{\partial \beta}.$$

The speed of $z(\beta)$ gets infinite.

The statement on the shape of $\tilde{r}_\nu(z)$ is proved by multiplying the equation for $\tilde{u}$ by α^2:

$$\begin{aligned} (\alpha z)(\alpha u) + (\alpha\nu)\Delta(\alpha u) - \nabla(\alpha^2 p) &= 0, \\ (\alpha u)_n|_\partial &= -(\alpha z)\Phi_0, \\ (\alpha^2 p - 2(\alpha\nu) S^n_{\alpha u})|_\partial &= \alpha^2 \tilde{r}_\nu(z). \end{aligned}$$

By definition of $\tilde{r}$ the last line coincides with $\tilde{r}_{\alpha\nu}(\alpha z)$. □

6. A Hopf bifurcation for liquid drops. In this section we show how our analysis can be used to study the effect of an exterior force. In the previous section, we achieved a complete picture of the spectrum of $\mathcal{L}$. On a fixed subspace X_k, the spectrum consists of a countable number of eigenvalues that correspond to eigenvalues of the Stokes operator and two additional eigenvalues (interfacial eigenvalues) that are real for small surface tension and nonreal for large surface tension.

Starting from this situation, a Hopf bifurcation can occur if an exterior force moves the additional eigenvalues across the imaginary axis. We prove this behavior in the case of a force that preserves symmetry. A more physical force will not preserve symmetry; we study that case in section 7.

In this section we assume that the force acts on the surface and that its strength depends linearly on the position and the speed of the boundary. The symmetric force has only two parts; the one proportional to η acts as the surface tension, and we restrict our analysis to a force proportional to $\partial_t \eta(x,t) = u_n|_\partial$. We introduce the real number λ for its strength and have the boundary condition

$$(p - \nu S_u^n)|_\partial + \beta \underline{\Delta} \eta = \lambda u_n|_\partial.$$

We write the linear equations again as

$$\frac{d}{dt} x + \mathcal{L}_\lambda x = 0, \qquad x \in \tilde{X},$$

now with the operator

$$\mathcal{L}_\lambda \begin{pmatrix} u \\ \eta \end{pmatrix} := \begin{pmatrix} -\nu \Delta u + \nabla \mathcal{H}(2\nu S_u^n) - \nabla \mathcal{H}(\beta \underline{\Delta} \eta) - \nabla \mathcal{H}(\lambda u_n|_\partial) \\ -u_n|_\partial \end{pmatrix}.$$

This operator is a lower order perturbation of $\mathcal{L}$; its spectrum consists of eigenvalues and we have the local existence results as before. The following analogue of Lemma 2.4 holds.

LEMMA 6.1. *Let $(u, \eta) \in \tilde{X}^2$ be an eigenvector of $\mathcal{L}_\lambda$ with eigenvalue μ. Then*

$$\begin{aligned} \operatorname{Re}(\mu) \|(u,\eta)\|_E^2 &= 2\nu \left(\int_B |S_u|^2 \right) - \lambda |\mu|^2 \|\eta\|^2_{L^2(S^2)}, \\ \operatorname{Im}(\mu) \|(u,\eta)\|_E^2 &= 2\beta \operatorname{Im} \left(\int_S u_n|_\partial \underline{\Delta} \overline{\eta} \right). \end{aligned} \tag{33}$$

In the case of nonreal eigenvalues, $\operatorname{Im}(\mu) \neq 0$*, the following energy equality holds:*

$$\|u\|_E^2 = \|\eta\|_E^2 = \frac{1}{2} \|(u,\eta)\|_E^2. \tag{34}$$

Proof. This lemma is proved as Lemma 2.4. □

We again want to get a global picture of the position of eigenvalues, now in dependence of the parameter λ. There are two important differences from the previous section:

— the eigenvalues may have a negative real part and

— the energy equality for eigenvectors implies that nonreal eigenvalues are bounded independent of λ.

For any $z \in \mathbb{R}_+ \setminus \{\kappa_j | j \in \mathbb{N}\}$ we have defined the function $\tilde{u}(z)$. The pair $(\tilde{u}(z), \Phi_0)$ is an eigenfunction of $\mathcal{L}_{\tilde{\lambda}}$, with $\tilde{\lambda}(z)$ defined by

$$\tilde{r}(z)\Phi_0 = (\tilde{p}(z) - 2\nu S^n_{\tilde{u}(z)})|_\partial = \beta \underline{\Lambda_k} \Phi_0 + \tilde{\lambda}(z) z \Phi_0. \tag{35}$$

PROPOSITION 6.2 (properties of $\tilde{\lambda}(z)$). *There exists a constant $\lambda_0 > 0$ such that*

$$|\lambda| > \lambda_0 \Rightarrow \textit{all eigenvalues of } \mathcal{L}_\lambda \textit{are real}, \tag{36}$$

$$\tilde{\lambda}(z) \to -\infty \ for \ \mathbb{R} \ni z \to 0, \tag{37}$$

$$\tilde{\lambda}(z) \to -\infty \ for \ \mathbb{R} \ni z \nearrow \kappa_j, \tag{38}$$

$$\tilde{\lambda}(z) \to +\infty \ for \ \mathbb{R} \ni z \searrow \kappa_j. \tag{39}$$

Proof. Equation (33) implies, for nonreal eigenvalues, $|\lambda| \to \infty \Rightarrow \int |S_u|^2 \to \infty$ or $|\mu| \to \infty$ or $|\mu| \to 0$. Therefore, $|\mu| \to \infty$ or $|\mu| \to 0$. (25) implies $\|u\| \to \infty$ or $\|u\| \to 0$. This contradicts (34). (37)–(39) follow from $\tilde{\lambda}(z) = \frac{1}{z}(\tilde{r}(z) - \beta \underline{\Lambda_k})$. □

THEOREM 6.3 (the spectrum of $\mathcal{L}$ in dependence of λ). *For $\lambda < -\lambda_0$, all eigenvalues of $\mathcal{L}_\lambda$ are real. Denoting the ordered sequence of them by $(\mu_j(\lambda))_{j\in\mathbb{N}}$, they satisfy*

$$\begin{aligned}
&0 < \mu_0(\lambda) < \mu_1(\lambda) < \kappa_0, \qquad \kappa_j < \mu_{j+2} < \kappa_{j+1},\\
&\mu_0(\lambda) \searrow 0 \textit{ for } \lambda \to -\infty \quad \textit{and}\\
&\mu_{j+2}(\lambda) \nearrow \kappa_{j+1} \textit{ for } \lambda \to -\infty.
\end{aligned}$$

For $\lambda > \lambda_0$ the ordered eigenvalues satisfy

$$\begin{aligned}
&\mu_0(\lambda), \mu_1(\lambda) < 0, \qquad \kappa_j < \mu_{j+2} < \kappa_{j+1},\\
&\mu_{j+2}(\lambda) \searrow \kappa_j \textit{ for } \lambda \to \infty.
\end{aligned}$$

In a point $\bar{\lambda} \in [-\lambda_0, \lambda_0]$, a pair of conjugate complex eigenvalues crosses the imaginary axis transversally. The imaginary axis can be crossed only from right to left.

Proof. Proposition 6.2 implies the assertion for the position of positive real eigenvalues for $|\lambda| \to \infty$. We have to prove the existence of the pair of negative eigenvalues. We do this by counting eigenvalues. (25), together with the energy equality (34), implies that nonreal eigenvalues are bounded independent of λ. Let κ_J be larger than this bound. We restrict ourselves to the $J+2$ eigenvalues of $\mathcal{L}_{-\lambda_0}$ with norm smaller than κ_J.

For $\lambda \to +\infty$ (and, in particular, $\lambda > \lambda_0$) there are J eigenvalues with positive real part and norm less than κ_J. The two remaining eigenvalues must be negative.

To count the eigenvalues, we used the fact that geometric and algebraic multiplicity coincide for $|\lambda| \to \infty$. We now prove this fact.

Assumption. There exist normed functions (u, η), (v, σ) satisfying

$$\mathcal{L}\begin{pmatrix} u \\ \eta \end{pmatrix} = \mu \begin{pmatrix} u \\ \eta \end{pmatrix}, \ \mathcal{L}\begin{pmatrix} v \\ \sigma \end{pmatrix} = \mu \begin{pmatrix} v \\ \sigma \end{pmatrix} + \alpha \begin{pmatrix} u \\ \eta \end{pmatrix}, \quad v \perp u.$$

Let p denote the pressure function corresponding to u. We know $(u, \eta) \to (u_j, 0)$ for $\lambda \to \infty$. We define $(v_0, \sigma_0) = \lim_{\lambda\to\infty}(v, \sigma)$. It holds that

$$-\nu\Delta v + \nabla\mathcal{H}(2\nu S_v^n) - \nabla\mathcal{H}(\beta\underline{\Delta}\sigma + \lambda v_n|_\partial) = \mu v + \alpha u, \tag{40}$$

$$-v_n = \mu\sigma + \alpha\eta. \tag{41}$$

To prove that α is bounded we multiply (40) by u_j and integrate to get

$$\begin{aligned}
\alpha \langle u, u_j\rangle &= \langle -\mu v - \nu\Delta v + \nabla\mathcal{H}(2\nu S_v^n) - \nabla\mathcal{H}(\beta\underline{\Delta}\sigma + \lambda v_n|_\partial), u_j\rangle\\
&= \langle v, (A_N - \mu)u_j\rangle + \int_{S^2} v_n(p_j - 2\nu S_{u_j}^n)\\
&= (\kappa_j - \mu)\langle v, u_j\rangle + \int_{S^2} v_n(p_j - 2\nu S_{u_j}^n).
\end{aligned}$$

Using $\eta \to 0$, this implies that α is bounded. Equation (40) implies $(v_0)_n = 0$, and we get, from equation (41), $\sigma_0 = 0$. With $\alpha_0 := \lim \alpha$,

$$(A_N - \kappa_j)v_0 = \alpha_0 u_j,$$

a contradiction to the simplicity of the Stokes eigenvalues.

We now derive explicit equations for the velocity of eigenvalues. We consider a differentiable family of eigenvalues $\mu(\lambda) \in \mathbb{C} \setminus \mathbb{R}$ with eigenfunctions $(u(\lambda), \Phi_0)$.

We first need an equation for $\mathrm{Im}\langle u, \partial_\lambda u\rangle$. We multiply the eigenvalue equation by $\partial_\lambda u$ to get, with the help of $\partial_\lambda u_n|_\partial = -\partial_\lambda \mu(\lambda)\Phi_0$,

$$\langle \mu(\lambda)u(\lambda), \partial_\lambda u(\lambda)\rangle = 2\nu \int_{B^3} S_u : S_{\partial_\lambda \bar{u}} - (\beta \underline{\Lambda_k} + \lambda\mu(\lambda))\partial_\lambda \bar{\mu}(\lambda)\|\Phi_0\|_{L^2}^2.$$

Taking the real part and using Lemma 6.1 yields

$$\begin{aligned}
\mathrm{Re}\,\langle \mu u, \partial_\lambda u\rangle &= \frac{1}{2}\partial_\lambda\left(2\nu\int_{B^3} S_{u(\lambda)} : S_{\bar{u}(\lambda)}\right)\\
&\quad -\beta\underline{\Lambda_k}\|\Phi_0\|_{L^2}^2 \mathrm{Re}(\partial_\lambda \mu(\lambda)) - \lambda\frac{1}{2}\partial_\lambda|\mu(\lambda)|^2\|\Phi_0\|_{L^2}^2\\
&= \frac{1}{2}\partial_\lambda\{\mathrm{Re}(\mu(\lambda))2\beta\underline{\Lambda_k}\|\Phi_0\|_{L^2}^2 + \lambda|\mu(\lambda)|^2\|\Phi_0\|_{L^2}^2\}\\
&\quad -\beta\underline{\Lambda_k}\|\Phi_0\|_{L^2}^2 \mathrm{Re}(\partial_\lambda \mu(\lambda)) - \lambda\frac{1}{2}\partial_\lambda|\mu(\lambda)|^2\|\Phi_0\|_{L^2}^2\\
&= \frac{1}{2}|\mu(\lambda)|^2\|\Phi_0\|_{L^2}^2.
\end{aligned}$$

Using $\mathrm{Re}\,\langle u, \partial_\lambda u\rangle = 0$, we arrive at

$$\mathrm{Im}\,\langle u, \partial_\lambda u\rangle\,\mathrm{Im}(\mu) = -\frac{1}{2}|\mu(\lambda)|^2\|\Phi_0\|_{L^2}^2. \tag{42}$$

Now we differentiate the eigenvalue equation with respect to λ, multiply with $\partial_\lambda u$, and take the imaginary part:

$$\begin{aligned}
0 &= \mathrm{Im}\,\langle \partial_\lambda(-\mu u), \partial_\lambda u\rangle + \mathrm{Im}\,\langle -\nu\Delta\partial_\lambda u + \nabla\partial_\lambda p, \partial_\lambda u\rangle\\
&= -\mathrm{Re}(\partial_\lambda\mu)\mathrm{Im}\,\langle u, \partial_\lambda u\rangle - \mathrm{Im}(\mu)\|\partial_\lambda u\|^2 - \mathrm{Im}\{\partial_\lambda(\beta\underline{\Lambda_k} + \lambda\mu)\partial_\lambda\bar{\mu}\}\|\Phi_0\|_{L^2}^2.
\end{aligned}$$

Multiplying with $2\mathrm{Im}(\mu)$ and inserting (42) yields

$$\mathrm{Re}(\partial_\lambda\mu)|\mu|^2\|\Phi_0\|_{L^2}^2 - 2|\mathrm{Im}(\mu)|^2\|\partial_\lambda u\|^2 = 2\mathrm{Im}(\mu)\mathrm{Im}(\mu\partial_\lambda\bar{\mu})\|\Phi_0\|_{L^2}^2.$$

On the imaginary axis, $\mathrm{Re}(\mu) = 0$, this formula simplifies to

$$-\mathrm{Re}(\partial_\lambda\mu)\|\Phi_0\|_{L^2}^2 = 2\|\partial_\lambda u\|^2.$$

It proves transversality and the direction of the crossing. □

The above results lead to a Hopf bifurcation. Due to the restricted regularity, one has to avoid the implicit function theorem in the proof and use degree theory. It allows us to replace the assumption of transversality by the following. The eigenvalue cannot follow the imaginary axis, i.e.,

$$\mu(\bar{\lambda}) \in i\mathbb{R} \Rightarrow \forall\varepsilon > 0 : \mathrm{Re}(\mu(\bar{\lambda} - \varepsilon)) > 0, \qquad \mathrm{Re}(\mu(\bar{\lambda} + \varepsilon)) < 0. \tag{43}$$

Theorem 6.4. *For fixed wave number k_0, there exists a critical value for λ such that a pair of eigenvalues $\mu^\pm$ of $\mathcal{L}_\lambda$ are purely imaginary. Assume that there is no resonance, i.e., the eigenvalues for different k are no integer multiples of μ^+. Then* *a Hopf bifurcation occurs and there exists a continuous branch of $O(2)$-symmetric, periodic solutions of the nonlinear equations.*

Proof. In Proposition 4.3 we parametrized small initial values for the nonlinear equation over $U = B_\varepsilon(0) \subset D_{\mathcal{L},\alpha}$ with a mapping $\zeta : U \to X^4$. We want to restrict ourselves to functions of the prescribed symmetry group $\Gamma \simeq O(2)$. We parametrize small compatible initial values with symmetry Γ over $V := U \cap Fix(\Gamma)$ with a map $\zeta_\Gamma : V \to Fix(\Gamma) \subset X^4$ and consider the flow

$$\Phi : V \times \mathbb{R} \times \mathbb{R} \ni (z, t, \lambda) \mapsto \zeta_\Gamma^{-1} x(t) - z \in V.$$

Here $x(t)$ is the solution of the nonlinear equation with parameter λ to the initial value $\zeta_\Gamma(z)$. We used the fact that the nonlinear equation preserves the Γ-symmetry. We want to solve $\Phi(z, t, \lambda) = z$ with nontrivial z. The linearization of Φ in $(0, \bar{t} = \frac{2\pi}{\mathrm{Im}(\mu_0(\bar{\lambda}))}, \bar{\lambda})$ is

$$D_z\Phi : V \to V, \quad z \mapsto e^{-\mathcal{L}_{\bar{\lambda}}\bar{t}} z,$$

and the kernel of $D_z\Phi(., \bar{t}, \bar{\lambda}) - id$ is two-dimensional. One can perform a Liapunov–Schmidt reduction and solve the bifurcation equation with degree theory. □

7. Water waves generated by wind.

We analyze a simple two-dimensional model for a wind-generated instability of a water surface. The wind changes the pressure along the surface. We assume that the pressure profile follows the sinusoidal profile of the surface and is shifted by an angle ϕ. Measurements of Elliott [7] justify this assumption and give the value of 135° for ϕ.

We mention the two major simplifications of this model: it neglects the tangential stress, and we linearize about the zero-solution instead of assuming an underlying shear flow. The method could be extended, but the desirable further development would be a two-phase model to explore the dependence of the force of the wave number. For qualitative studies we refer to [12] and references therein.

With a surface elevation η we write the additional pressure as $e^{i\phi}\beta^*\eta$. This is equivalent to saying that we treat the complex surface tension $\beta + e^{i\phi}\beta^*\Lambda_k^{-1}$. We fix the direction of the wind by setting $0 < \phi < \pi$. The eigenvalue equations are

$$\lambda u + \nu\Delta u - \nabla p = 0, \tag{44}$$

$$\nabla \cdot u = 0, \tag{45}$$

$$u_n|_\partial = -\lambda\eta, \tag{46}$$

$$(p - 2\nu S_u^n)|_\partial = -\beta\underline{\Delta}\eta + e^{i\phi}\beta^*\eta. \tag{47}$$

Every eigenspace is (at least) two-dimensional due to the equivariance under translations. Choosing $\eta(x) = \Phi_0(x) = e^{ikx}$, we select one of the eigenfunctions. On the other hand, a positive β^* destroys the symmetry of reflections $x \mapsto -x$. The conjugate complex of an eigenvalue need not be an eigenvalue for the same η.

Lemma 7.1. *For positive β^* there is precisely one eigenvalue with positive imaginary part. The other eigenvalues are below the real axis and have a positive real part.*

Proof. As in section 5 we can write the eigenvalue equation as

$$\tilde{r}(z) = \underline{\Lambda_k}\beta + e^{i\phi}\beta^*. \tag{48}$$

The proof of the lemma is based on the study of $\tilde{r}$ in Proposition 5.3.

For $\beta^* = 0$ we consider two cases.

Case 1. The surface tension is above its critical value. Then there are two nonreal eigenvalues with multiplicity 2. Due to the reflection symmetry, they can both be represented with an eigenfunction with $\eta(x) = \Phi_0(x)$.

The function $\tilde{r} : \mathbb{R} \to \mathbb{R}$ was shown to have a negative derivative in the real eigenvalues. Since the function $\tilde{r}$ is analytic in a neighborhood of the eigenvalue and β^* acts like an imaginary component of the surface tension, we conclude that for small, positive β^* the eigenvalues get a negative imaginary part.

Case 2. The surface tension is below its critical value. Then the derivative of $\tilde{r} : \mathbb{R} \to \mathbb{R}$ is positive in the first eigenvalue, negative in all other eigenvalues. The same reasoning as before proves that for small positive β^* the first eigenvalue gets a positive imaginary part, and the other eigenvalues get a negative imaginary part.

In both cases a return to the real axis is not possible for finite β^* since real eigenvalues correspond to real surface tension or $\beta^* = 0$.

We derive an equation for λ by testing the eigenvalue equation with (u, η) in the energy space and using the integration by parts:

$$\begin{aligned}\int (-\nu \Delta u + \nabla p)\bar{u} &= 2\nu \int |S_u|^2 + \int_\partial (p - 2\nu S_u^n)\bar{u}_n \\ &= 2\nu \int |S_u|^2 - \int_\partial (-\beta \underline{\Delta} \eta + e^{i\phi}\beta^* \eta)\bar{\lambda}\bar{\eta}.\end{aligned}$$

In analogy to Lemma 2.4, we get

$$\lambda\{\|u\|^2 + \|\eta\|_E^2\} = 2\nu \int |S_u|^2 + 2i\mathrm{Im}\lambda \|\eta\|_E^2 - e^{i\phi}\beta^*\bar{\lambda}\|\eta\|_{L^2}^2. \tag{49}$$

Taking the real part of this equation, we get

$$\mathrm{Re}\lambda = 0 \Rightarrow 0 < \mathrm{Re}(e^{-i\phi}\lambda) \Rightarrow \mathrm{Im}(\lambda) > 0.$$

This proves that the eigenvalues with negative imaginary part cannot cross the imaginary axis. □

The only eigenvalue that can create an instability is the one with positive imaginary part, further denoted by λ^+. We turn to an analysis of this eigenvalue.

We know that for eigenvalues λ in a compact subset of $\mathbb{C} - \{\kappa_j | j \in \mathbb{N}\}$ the corresponding values of $\beta^* = e^{-i\phi}(\tilde{r}(\lambda) - \Lambda_k \beta)$ are finite. Therefore, for $\beta^* \to \infty$,

$$\begin{aligned}\exists j :&\quad \lambda^+(\beta^*) \to \kappa_j \\ \text{or}&\quad |\lambda^+(\beta^*)| \to \infty.\end{aligned}$$

We take the imaginary part of (49) and get

$$\mathrm{Im}\lambda^+\{\|u\|^2 - \|\eta\|_E^2\} = -\beta^*\mathrm{Im}(e^{i\phi}\bar{\lambda}^+)\|\eta\|_{L^2}^2. \tag{50}$$

Assume that $\mathrm{Im}(\lambda^+)$ stays bounded. Then we know $|\mathrm{Re}(\lambda^+)| \to \infty$. The left-hand side of (50) is bounded from below. We conclude that $\mathrm{Re}(\lambda^+) \to -\infty$.

The eigenvalue must cross the imaginary axis, by the following lemma.

Lemma 7.2. *An eigenvalue* λ *with* $|\mathrm{Im}\lambda| \to \infty$ *or* $\mathrm{Re}(\lambda) \to -\infty$ *as* $\beta^* \to \infty$ *satisfies*

$$\arg(\lambda) \to \frac{\phi \pm \pi}{2} \quad \textit{for } \beta^* \to \infty.$$

Proof. To prove this proposition we use the explicit formulas for eigenfunctions. We introduce

$$v = u - \frac{1}{\lambda}\nabla p, \tag{51}$$

which solves $\lambda v + \nu\Delta v = 0$. With constants $A = (A_1, A_2), B = (B_1, B_2), P$, and Q we write

$$\begin{aligned} p(x,y) &= Pe^{ky}e^{ikx} + Qe^{-ky}e^{ikx},\\ v(x,y) &= Ae^{\mu y}e^{ikx} + Be^{-\mu y}e^{ikx},\\ \mu^2 &= k^2 - \frac{\lambda}{\nu}, \end{aligned}$$

where we take μ as the root with positive real part, $\mathrm{Re}\mu \to \infty$.

The incompressibility reads

$$ikA_1 + \mu A_2 = 0, \qquad ikB_1 - \mu B_2 = 0. \tag{52}$$

We get an equation for P and Q if we construct

$$\begin{aligned} 0 &= iku_1(-h) + \mu u_2(-h)\\ &= \frac{k^2}{\lambda}(Pe^{-kh} + Qe^{kh}) - \frac{k\mu}{\lambda}(Pe^{-kh} - Qe^{kh}), \end{aligned}$$

which proves that $|Q/P|$ is bounded, and then

$$\left(1 - \frac{Q}{P}e^{2kh}\right) = \frac{1}{P}(P - Qe^{2kh}) \to 0. \tag{53}$$

Using $u(x,-h) = 0$, this implies

$$\frac{B_1}{|A_1| + |P|} \to 0 \qquad \text{and} \qquad \frac{B_2}{|A_2| + |P|} \to 0 \qquad \text{exponentially in } \mu.$$

We use the boundary condition of vanishing tangential stress,

$$0 = A_1\mu - B_1\mu + A_2 ik + B_2 ik + \frac{2ik^2}{\lambda}(P - Q),$$

to conclude that

$$\frac{A_1\mu\lambda}{ik^2(P - Q)} \to -1. \tag{54}$$

The boundary condition for the normal stress reads

$$0 = (P + Q) - 2\nu(A_2 - B_2)\mu - \frac{2\nu}{\lambda}k^2(P + Q) - \beta\Lambda_k - e^{i\phi}\beta^*.$$

Collecting the dominant terms yields

$$\frac{P + Q}{\beta^*} \to e^{i\phi}. \tag{55}$$

The kinematic boundary equation is

$$A_2 + B_2 + \frac{k}{\lambda}(P - Q) = -\lambda.$$

Using (54), this implies

$$\frac{k}{\lambda^2}(P - Q) \to -1. \tag{56}$$

Combining (53), (55), and (56) gives

$$\frac{\lambda^2}{e^{i\phi}\beta^*} \to -k\frac{1 - e^{-2kh}}{1 + e^{-2kh}}. \qquad \square \tag{57}$$

Remark 7.3. *The basic idea in the above proof was to show that the dominating term in the stress equation is the pressure. In this sense the following formal calculation for infinite height and vanishing viscosity is justified:*

$$\Delta p = 0, \quad p|_\partial = e^{i\phi}\beta^*\eta, \quad u = \frac{1}{\lambda}\nabla p,$$

$$-\lambda\eta = u_n|_\partial = \frac{1}{\lambda}\partial_n p = \frac{k}{\lambda}e^{i\phi}\beta^*\eta.$$

Remark 7.4. *To reduce the formula to the case of vanishing wind but variable surface tension T, one can insert $\beta^* = \Lambda_k T = k^2 T$ to get the asymptotic formula*

$$(-i\lambda)^2 = Tk^3\frac{1 - e^{-2kh}}{1 + e^{-2kh}}. \tag{58}$$

Formulas that include the viscous effect and give a similar expression for the real part of λ can be found in [11].

In this example it is not easy to prove transversality of the crossing. On the other hand, we can easily verify property (43). If the eigenvalue followed the imaginary axis then $e^{-i\phi}\tilde{r}(z)$ were real on an interval of the imaginary axis. By analyticity, it would be real on the whole of the imaginary axis, which contradicts the last lemma.

With the proof of the last section we arrived at the following theorem.

Theorem 7.5. *Assuming the nonresonance condition, for some critical value of β^* a Hopf bifurcation occurs and periodic solutions of the nonlinear evolution equations exist.*

We remark that the constructed solution is a propagating wave but not necessarily a travelling wave.

REFERENCES

[1] S. Agmon, A. Douglis, and N. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[2] J. T. Beale, *Large time regularity of viscous surface waves*, Arch. Rational Mech. Anal., 84 (1984), pp. 307–352.

[3] E. Becker, W. J. Hiller, and T. A. Kowalewski, *Nonlinear dynamics of viscous droplets*, J. Fluid Mech., 258 (1994), pp. 191–216.

[4] J. Bemelmans, *Gleichgewichtsfiguren zäher Flüssigkeiten mit Oberflächenspannung*, Analysis, 1 (1981), pp. 241–282.

[5] M. G. Crandall and P. H. Rabinowitz, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rational Mech. Anal., 67 (1978), pp. 53–72.

[6] G. da Prato, *Abstract differential equations, maximal regularity, and linearization*, Nonlinear Functional Analysis and Its Applications, Proc. Symp. in Pure Math., 45/1 (1986), pp. 359–370.

[7] J. A. Elliott, *Microscale pressure fluctuations near waves generated by wind*, J. Fluid Mech., 54 (1972), pp. 427–448.

[8] M. Golubitsky, I. Stewart, and D. Schaeffer, *Singularities and Groups in Bifurcation Theory* 2, Applied Mathematical Sciences 69, Springer-Verlag, Berlin, 1988.

[9] H. Koch, *On a Fully Nonlinear Mixed Parabolic Problem with Oblique Boundary Condition*, Preprintserie des IWR 95-20, Heidelberg, Germany, 1995.

[10] H. Lamb, *Lehrbuch der Hydrodynamik*, Lehrbuch der mathematischen Naturwissenschaften, 26, Teubner, Leipzig, 1931.

[11] T. S. Lundgren and N. N. Mansour, *Oscillations of drops in zero gravity with weak viscous effects*, J. Fluid Mech., 194 (1988), pp. 479–510.

[12] L. C. Morland and P. N. Saffman, *Effect of wind profile on the instability of wind blowing over water*, J. Fluid Mech., 252 (1993), pp. 383–398.

[13] M. Renardy and D. D. Joseph, *Hopf bifurcation in two component flow*, SIAM J. Math. Anal., 17 (1986), pp. 894–910.

[14] B. Schweizer, *Oscillatory behavior of liquid drops*, Preprintserie des IWR 95-34, Heidelberg, Germany, 1995.

[15] V. A. Solonnikov, *The solvability of the problem concerning the evolution of an isolated volume of viscous incompressible capillary fluid*, J. Soviet Math., 32 (1986), pp. 223–228.

[16] V. A. Solonnikov and V. Skadilov, *On a boundary value problem for a stationary system of the Navier-Stokes equations*, Proc. Steklov Inst. Math., 125 (1973), pp. 186–199.

[17] A. Wagner, *Stationary Marangoni Convection: A Free Boundary Problem for the Navier Stokes Equations*, Preprintserie des IWR 94-41, Heidelberg, Germany, 1994.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1158–1172, September 1997

A SEMIGROUP APPROACH TO FRAGMENTATION MODELS*

D. J. McLAUGHLIN†, W. LAMB†, AND A. C. McBRIDE†

Abstract. An initial-value problem modelling fragmentation processes, where particles split into two or more pieces, is studied using the theory of linear semigroups. The existence and uniqueness of nonnegative, mass-conserving solutions are established.

Key words. semigroup theory, fragmentation equation

AMS subject classifications. 35F10, 45K05, 47D06, 47N20, 80A30

PII. S0036141095291701

1. Introduction. There are two main descriptions of coagulation and fragmentation processes, a "continuous" integral version and a "discrete" summation version. In the continuous version it is assumed that the number of particles is large enough to justify the use of a density function $u(x,t)$; $u(x,t)dx$ is then the average number of particles with mass in the interval $(x, x+dx)$ at time t. (This average and all other averages are calculated with respect to a unit volume.)

This paper is the first in a series of planned papers reporting our investigation of the continuous coagulation and multiple-fragmentation equation, namely,

$$\begin{aligned}\frac{\partial}{\partial t}u(x,t)=&\frac{1}{2}\int_0^x K(x-y,y,t)u(x-y,t)u(y,t)\,dy-u(x,t)\int_0^\infty K(x,y,t)u(y,t)\,dy\\ &+\int_x^\infty \gamma(y,x,t)u(y,t)\,dy-u(x,t)\int_0^x \frac{y}{x}\gamma(x,y,t)\,dy, \quad \text{a.e. } x>0,\ t>0,\end{aligned}\tag{1}$$

via semigroup and evolution system techniques.

In the first instance we shall deal simply with the pure fragmentation equation ($K \equiv 0$) for time-independent kernels $\gamma(x,y,t) = \gamma(x,y)$, for all $t > 0$. Therefore, we consider the equation

$$\frac{\partial}{\partial t}u(x,t) = \int_x^\infty \gamma(y,x)u(y,t)\,dy-u(x,t)\int_0^x \frac{y}{x}\gamma(x,y)\,dy, \quad \text{a.e. } x>0,\ t>0.\tag{2}$$

Future papers will extend this work to the coagulation and fragmentation equation (1) for both time-independent and time-dependent kernels. We begin by discussing the interpretation of the terms on the right-hand side of (2) and delay comment on the terms involved in (1) until the need arises.

Equation (2) models fragmentation processes in which particles may split into more than two pieces. Hence we refer to the latter as the multiple-fragmentation equation. When all of the fragmenting particles in the system each produce only two particles, the binary fragmentation equation,

$$\frac{\partial}{\partial t}u(x,t) = \int_0^\infty F(x,y)u(x+y,t)\,dy-\frac{1}{2}u(x,t)\int_0^x F(x-y,y)\,dy,\tag{3}$$

[*] Received by the editors September 8, 1995; accepted for publication (in revised form) May 1, 1996. The research of the first author was supported by the Engineering and Physical Sciences Research Council of the United Kingdom.

http://www.siam.org/journals/sima/28-5/29170.html

[†] Department of Mathematics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK (d.j.mclaughlin@strath.ac.uk, w.lamb@strath.ac.uk, a.c.mcbride@strath.ac.uk).

provides an alternative description. As expected, (3) is a special case of (2), obtained by setting

$$\gamma(x,y) = F(x-y,y), \quad 0 \le y \le x < \infty, \tag{4}$$

where $F(x,y)$ is symmetric; i.e., $F(x,y) = F(y,x)$ for all $x, y \ge 0$. Using the substitution $y' = x - y$ and the symmetry of F, it is not difficult to show that

$$\int_0^x \frac{y}{x} F(x-y,y)\,dy = \frac{1}{2}\int_0^x F(x-y,y)\,dy.$$

We develop our theory in terms of the multiple-fragmentation model (2) but note that, as a consequence of the foregoing discussion, our results immediately carry over to the binary-fragmentation model (3).

The multiple-fragmentation kernel $\gamma(x,y)$, $0 \le y \le x < \infty$, is the formation rate of particles of mass y due to the fragmentation of particles of mass x. We assume that $\gamma(x,y)u(x,t)\,dx\,dy\,dt$ is the average number of particles of mass in the range $(y, y+dy)$ created from the break-up of particles of mass in $(x, x+dx)$ during the time interval $(t, t+dt)$. The interpretation of the binary-fragmentation kernel $F(x,y)$ is slightly different. The binary-fragmentation kernel $F(x,y)$ is the rate at which particles of mass $x+y$ fragment to produce particles of mass x and of mass y. However, we can interpret the binary-fragmentation kernel $F(x-y,y)$ in terms of the multiple-fragmentation kernel $\gamma(x,y)$ using (4).

Note that $a : (0,\infty) \to [0,\infty)$ defined by

$$a(x) = \int_0^x \frac{y}{x}\gamma(x,y)\,dy, \qquad x > 0, \tag{5}$$

is the overall rate of break-up of an x-particle. The factor y/x is introduced so that each resulting fragment is counted once only. By writing $\gamma(x,y) = a(x)b(y,x)$ we obtain precisely the derivation of the fragmentation terms by Ziff and McGrady [20], where $a(x)$ is the overall break-up rate and $b(y,x)$ is the distribution of products formed from a particle of mass x splitting.

Thus the terms on the right-hand side of (2) represent, respectively, a gain in particles of mass x as a result of the break-up of larger particles of mass y $(x \le y < \infty)$ and a loss of particles of mass x because they have broken up into smaller pieces of mass y $(0 \le y \le x)$. Equation (2) is formulated so that the total mass of the particles is formally conserved. Thus $\frac{\partial u}{\partial t}(x,t)$ equals the sum of the two terms described above.

Many authors have already contributed to the study of existence and uniqueness theorems for the various coagulation and fragmentation models. Aizenman and Bak [1] consider the case when both K and F are constant. The semigroup and fixed point mapping techniques used in [1] form the basis of our work. Melzak [13, 14] deals with certain bounded time-independent kernels by assuming solutions are of the form $u(x,t) = \Sigma_{k=0}^{\infty} a_k(x)t^k$ and extends the results to time-dependent kernels via Cauchy–Peano approximations. The pure coagulation equation ($\gamma \equiv 0$ or $F \equiv 0$) and the pure fragmentation equation ($K \equiv 0$) have also received attention. Stewart [18, 19] deals simultaneously with unbounded coagulation and unbounded fragmentation kernels satisfying the growth conditions

$$\begin{aligned} K(x,y) &\le C_1[(1+x)^\alpha + (1+y)^\alpha] \quad (x,\, y > 0,\ 0 < \alpha < 1), \\ F(x,y) &\le C_2(1+x+y)^\beta \qquad (x,\, y > 0,\ 0 < \beta < 1). \end{aligned}$$

For physical reasons, solutions are required to satisfy

$$\int_0^\infty xu(x,t)\,dx = \int_0^\infty xf(x)\,dx \qquad \text{for all } t \ge 0,$$

where f is the initial mass distribution of particles. This corresponds to mass conservation. The semigroup approach that we adopt here will be seen to lead to unique, mass-conserving solutions; the possibility of adapting this approach to cater to other, non-mass-conserving, solutions is a topic for future research. Although we shall concentrate on models derived from (1) and use the theory of semigroups, other approaches have been adopted which rely on different formulations. One instance is [4], where Filippov discusses multiple fragmentation via the theory of Markov processes. The equations used involve expectations of masses, based on the probability of a particle of a certain size splitting. Filippov obtains explicit solutions in particular cases and also considers the possibility of mass loss or "disintegration."

2. A reformulation of the problem. To apply the theory of semigroups of linear operators we must recast (2), subject to the initial mass distribution

$$u(x,0) = f(x), \quad \text{a.e. } x > 0, \tag{6}$$

as an abstract Cauchy problem (ACP). For each fixed $t \geq 0$, we define a function $u_{(t)}\colon (0,\infty) \to \mathbf{R}$ of the "mass" variable x by $u_{(t)}(x) = u(x,t)$, for a.e. $x > 0,\ t \geq 0$. Hence we can define a vector-valued function $\tilde{u}$,

$$\tilde{u}(t) = u_{(t)}, \quad t \geq 0, \tag{7}$$

from $[0,\infty)$ into an appropriate class of functions X.

For our purposes X will be a Banach space so that $d\tilde{u}/dt$ can be interpreted as the strong derivative of the vector-valued function $\tilde{u}$. As we are interested in mass-conserving solutions, a natural space to work in is the $L_{1,-1}$-space of equivalence classes of measurable, real-valued functions ϕ such that

$$\|\phi\|_{1,-1} = \int_0^\infty x|\phi(x)|\,dx < \infty.$$

Note that this is a Banach space of type L [6, pp. 69–70] enabling us to reinterpret (2) as follows. The left-hand side, $\partial u/\partial t$, can be thought of as the derivative with respect to t of the function $\tilde{u} : [0,\infty) \to L_{1,-1}$ defined by (7). For fixed $t > 0$, we can write the right-hand side of (2) as $A\tilde{u}(t)$, where the operator $A : L_{1,-1} \supseteq D_{\max}(A) \to L_{1,-1}$ is defined on its maximal domain $D_{\max}(A)$ by

$$[A\phi](x) = \int_x^\infty \gamma(y,x)\phi(y)dy - \phi(x)\int_0^x \frac{y}{x}\gamma(x,y)\,dy, \quad \phi \in D_{\max}(A). \tag{8}$$

The initial condition (6) becomes $\tilde{u}(0) = f$.

Hence the pure fragmentation equation (2), subject to the initial mass distribution (6), can be recast as the ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= Au(t), \quad t > 0, \\ u(0) &= f. \end{aligned} \tag{9}$$

(For ease of notation we have omitted the tilde.)

As indicated, we follow closely the approach used in [1]. However, instead of working in the subspaces $\mathcal{L}_n$ $(n > 0)$ defined by

$$\mathcal{L}_n = \{\phi \in L_{1,-1}\colon \phi \equiv 0 \text{ on } [n,\infty)\}, \tag{10}$$

we make use of the projection operators P_n $(n > 0)$ on $L_{1,-1}$ defined by

$$(P_n\phi)(x) = \begin{cases} \phi(x), & 0 < x < n, \\ 0, & x \geq n. \end{cases} \tag{11}$$

Details of a subspace approach adopted in the preliminary stages of our work can be found in [10, 11, 12]. The results contained therein are summarized in [8].

Throughout this paper we assume that the fragmentation kernel $\gamma(x, y)$ satisfies the following hypotheses:

(H1) $\gamma(x, y)$ is a nonnegative function on $(0, \infty) \times (0, \infty)$;

(H2) $\gamma(x, y) = 0$ whenever $y > x$;

(H3) the function $a\colon (0, \infty) \to [0, \infty)$ defined by (5) is such that $a(x) \leq C_n$ for all $x \in (0, n]$, $n > 0$, where the sequence $\{c_n\}$ may be unbounded.

3. The truncated problem. The truncated problem consists of finding a solution to the truncated ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= A_n u(t), \quad t > 0, \\ u(0) &= f, \end{aligned} \tag{12}$$

where A_n is used to denote AP_n. Therefore, for $\phi \in L_{1,-1}$,

$$(A_n\phi)(x) = \begin{cases} \int_x^n \gamma(y, x)\phi(y)\,dy - \phi(x)\int_0^x \dfrac{y}{x}\gamma(x, y)\,dy, & 0 < x < n, \\ 0, & x \geq n. \end{cases}$$

We can write this as

$$(A_n\phi)(x) = \begin{cases} \int_x^n \gamma(y, x)\phi(y)\,dy - a(x)\phi(x), & 0 < x < n, \\ 0, & x \geq n, \end{cases} \tag{13}$$

where $a : (0, \infty) \to [0, \infty)$ is defined by (5).

LEMMA 3.1. *The operator A_n, given by* (13), *generates a C_0-semigroup $\{S_n(t)\}_{t\geq 0}$ on $L_{1,-1}$. Furthermore, for all $t \geq 0$,*

(i) $S_n(t) = I + \sum_{i=1}^{\infty} \frac{(tA)^i}{i!} P_n$;

(ii) *for $m > 0$ the space $\mathcal{L}_m$ defined by* (10) *is invariant under $S_n(t)$;*

(iii) *$S_n(t)\phi \geq 0$ whenever $\phi \geq 0$, i.e., the operators are nonnegative;*

(iv) *$\|S_n(t)\phi\|_{1,-1} = \|\phi\|_{1,-1}$, i.e., $\{S_n(t)\}_{t\geq 0}$ is a C_0-semigroup of isometries;*

(v) *$\int_0^\infty x[S_n(t)\phi](x)\,dx = \int_0^\infty x\phi(x)\,dx$, i.e., mass is conserved.*

Proof. Using Fubini's theorem we can show that for all $\phi \in L_{1,-1}$,

$$\|A_n\phi\|_{1,-1} = \int_0^\infty x|(A_n\phi)(x)|\,dx = \int_0^n x|(A_n\phi)(x)|\,dx \leq 2C_n\,\|\phi\|_{1,-1},$$

by (5) and (H3). Therefore A_n is a bounded operator on $L_{1,-1}$ and as such generates a uniformly (and hence strongly) continuous semigroup.

(i) The usual power series definition is used to define $S_n(t) = \exp(tA_n)$. By induction, $(A_n)^i = (AP_n)^i = A^i P_n$ for $i = 1, 2, \ldots$, from which the result follows.

(ii) We note that

$$A_n\phi = \begin{cases} A_m\phi, & \text{if } m \leq n, \\ A_n\phi, & \text{if } m > n, \end{cases}$$

and $A_n : L_{1,-1} \to \mathcal{L}_n$. The result follows from (i).

(iii) We can write $A_n\phi = B_{1,n}\phi + B_{2,n}\phi$ where

$$(B_{1,n}\phi)(x) = \begin{cases} \int_x^n \gamma(y,x)\phi(y)\,dy, & \text{if } x < n, \\ 0, & \text{if } x \geq n, \end{cases}$$

and

$$(B_{2,n}\phi)(x) = \begin{cases} -a(x)\phi(x), & \text{if } x < n, \\ 0, & \text{if } x \geq n. \end{cases} \tag{14}$$

From the calculations for $\|A_n\phi\|_{1,-1}$, we can show that the operators $B_{i,n}$ $(i = 1, 2)$ are bounded separately on $L_{1,-1}$ and hence generate C$_0$-semigroups on $L_{1,-1}$.

(a) The operator $B_{1,n}$ generates the C$_0$-semigroup, $T_{1,n}(t) = \exp(tB_{1,n})$, $t \geq 0$, where again the usual power series is used to define the exponential. Since $B_{1,n}$ is nonnegative, the semigroup $\{T_{1,n}(t)\}_{t\geq 0}$ is nonnegative. Moreover, $\{T_{1,n}(t)\}_{t\geq 0}$ is of class C$(1, \omega_{1,n})$, for some $\omega_{1,n} > 0$.

(b) Consider $B_{2,n}$ defined by (14) and let $\phi \in L_{1,-1}$. Then

$$(\lambda I - B_{2,n})\phi = 0 \Rightarrow \phi(x)[\lambda + a(x)] = 0 \qquad \text{for a.e. } x > 0.$$

Since $\gamma(x, y)$ is nonnegative, it follows that $a(x)$ is nonnegative and hence that $[\lambda + a(x)] \geq \lambda > 0$ for all $\lambda > 0$. We deduce that for all $\lambda > 0$, $(\lambda I - B_{2,n})$ is injective. To find the resolvent operator $R(\lambda, B_{2,n})$, let $\psi \in L_{1,-1}$. Then

$$\begin{aligned} (\lambda I - B_{2,n})\phi = \psi &\Rightarrow \lambda\phi(x) + a(x)\phi(x) = \psi(x) \quad \text{a.e.} \\ &\Rightarrow \phi(x) = \frac{\psi(x)}{[\lambda + a(x)]} \quad \text{a.e.} \end{aligned}$$

and ϕ belongs to $L_{1,-1}$ since $|\phi(x)| \leq \frac{1}{\lambda}|\psi(x)|$ for all $\lambda > 0$. Hence $B_{2,n}$ generates a contraction semigroup (that is, of class C$(1, 0)$), by the Hille–Yosida theorem. If $\psi \geq 0$, then $\phi(x) = R(\lambda, B_{2,n})\psi(x) \geq 0$ almost everywhere. Therefore, $R(\lambda, B_{2,n})$ is nonnegative and $B_{2,n}$, like $B_{1,n}$, generates a nonnegative semigroup, which we call $\{T_{2,n}(t)\}_{t\geq 0}$.

(c) From part (a) and the Hille–Yosida theorem, the resolvent set of $B_{1,n}$ contains the set $\{\lambda\colon \lambda > \omega_{1,n}\}$ and for all such λ, $\|R(\lambda, B_{1,n})\|_{1,-1} \leq 1/(\lambda - \omega_{1,n})$. Likewise, from (b), the resolvent set of $B_{2,n}$ contains the set $\{\lambda\colon \lambda > 0\}$ and for all such λ, $\|R(\lambda, B_{2,n})\|_{1,-1} \leq 1/\lambda$. By [15, Cor. 5.5, pp. 92–93], A_n generates a C$_0(1, \omega_{1,n})$-semigroup, $\{S_n(t)\}_{t\geq 0}$ say, where

$$[S_n(t)]\phi = \lim_{m\to\infty} \left[T_{1,n}\left(\frac{t}{m}\right) T_{2,n}\left(\frac{t}{m}\right)\right]^m \phi, \quad \text{for all } \phi \in L_{1,-1},$$

and the limit is uniform on bounded time intervals. The terms in the sequence are nonnegative since $T_{1,n}(t)$ and $T_{2,n}(t)$ are nonnegative. By [16, Cor. 5.11, p. 72] there exists a subsequence $\{m_k\}_{k=1}^{\infty}$ such that we have pointwise convergence almost everywhere. We deduce that the limit (and hence the semigroup generated by A_n) is nonnegative.

(iv) Assume $\phi \geq 0$ and $\phi \in \bigcup_m \mathcal{L}_m$. Then $\phi \in \mathcal{L}_m$ for some fixed $m > 0$.

Since $S_n(t)$ is nonnegative and $S_n(t)\phi \in \mathcal{L}_m$ by part (ii),

$$\begin{aligned} \frac{d}{dt}\|S_n(t)\phi\|_{1,-1} &= \frac{d}{dt}\int_0^m x[S_n(t)\phi](x)\,dx \\ &= \int_0^m x\frac{d}{dt}[S_n(t)\phi](x)\,dx \end{aligned}$$

$$
\begin{aligned}
&= \int_0^m x[A_n S_n(t)\phi](x)dx \\
&= \int_0^n x[A_n S_n(t)\phi](x)\,dx \quad (\text{since } A_n\colon L_{1,-1} \to \mathcal{L}_n) \\
&= \int_0^n \int_0^y \frac{x}{y}\gamma(y,x)\,dx\, y[S_n(t)\phi](y)dy - \int_0^n xa(x)[S_n(t)\phi](x)\,dx \\
&= 0 \quad (\text{by } (5)).
\end{aligned}
$$

It follows that $\|S_n(t)\phi\|_{1,-1} = \|\phi\|_{1,-1}$ for all $\phi \in \mathcal{L}_m$, $\phi \geq 0$. Since the operators $\{S_n(t)\}_{t\geq 0}$ are linear and nonnegative, we can use the decomposition of ϕ into positive and negative parts to extend this result to all $\phi \in \mathcal{L}_m$. Then, since $\bigcup_m \mathcal{L}_m$ is dense in $L_{1,-1}$, we can extend again to all $\phi \in L_{1,-1}$ by the continuity of $S_n(t)$.

(v) Since $S_n(t)$ is a nonnegative isometry, we can use positive and negative parts once more and hence remove the $|.|$-sign from the result of part (iv). □

Theorem 3.2. *The truncated problem* (12) *has a unique, strongly continuously differentiable, nonnegative, mass-conserving solution for all initial data $f \in L_{1,-1}$. The solution is given by $u(t) = S_n(t)f$, $t \geq 0$.*

Proof. This follows by Lemma 3.1 and [15, Thm. 1.3, pp. 102–103]. □

4. The limit semigroup.

In this section we show that the strong limit as $n \to \infty$ of the semigroups $\{S_n(t)\}_{t\geq 0}$, obtained in the previous section, exists and inherits the appropriate properties from $\{S_n(t)\}_{t\geq 0}$.

Lemma 4.1.

(i) *The operator $S(t)$ defined by*

$$S(t)\phi = \operatorname*{s-lim}_{n\to\infty} S_n(t)\phi, \qquad \phi \in L_{1,-1}, \tag{15}$$

exists for all $t \geq 0$ and $\{S(t)\}_{t\geq 0}$ forms a nonnegative, C_0-semigroup of isometries on $L_{1,-1}$.

(ii) *For each $t \geq 0$, the operator $S(t)$ satisfies*

$$\int_0^\infty x[S(t)\phi](x)dx = \int_0^\infty x\phi(x)dx, \quad \textit{for all } \phi \in L_{1,-1}.$$

(iii) *For all $t \geq 0$ and $m > 0$, $S(t)P_m\phi = S_m(t)\phi + P_m\phi - \phi$.*

(iv) *The spaces $\mathcal{L}_m$ $(m > 0)$ and $\bigcup_m \mathcal{L}_m$, where $\mathcal{L}_m$ is defined by* (10), *are invariant under $S(t)$.*

Proof. Let $\phi \in L_{1,-1}$.

(i) and (ii). Assume without loss of generality that $n \geq m$. Using Lemma 3.1(i), we can show that

$$S_n(t)P_m\phi = S_m(t)\phi + P_m\phi - \phi. \tag{16}$$

It is then not difficult to show that $\{S_n(t)\phi\}_{n>0}$ is a Cauchy sequence in $L_{1,-1}$ and hence the limit, which is uniform in t, exists, and $S(t)$ maps $L_{1,-1}$ into $L_{1,-1}$.

We can show that $S(t)$ inherits the properties of $S_n(t)$. For example, by definition,

$$S(t+s)\phi = \operatorname*{s-lim}_{n\to\infty} S_n(t+s)\phi = \operatorname*{s-lim}_{n\to\infty} S_n(t)S_n(s)\phi.$$

Consider

$$
\begin{aligned}
&\|S_n(t)S_n(s)\phi - S(t)S(s)\phi\|_{1,-1} \\
&\leq \|S_n(t)\{S_n(s)\phi - S(s)\phi\}\|_{1,-1} + \|\{S_n(t) - S(t)\}S(s)\phi\|_{1,-1} \\
&= \|\{S_n(s)\phi - S(s)\phi\}\|_{1,-1} + \|\{S_n(t) - S(t)\}\{S(s)\phi\}\|_{1,-1} \\
&\qquad\qquad (\text{since } \{S_n(t)\} \text{ is a semigroup of isometries on } L_{1,-1}) \\
&\to 0 \ (\text{by the definition of } S(t),\ t \geq 0).
\end{aligned}
$$

Hence $S(t+s)\phi = S(t)S(s)\phi$ for all $\phi \in L_{1,-1}$, and so $S(t+s) = S(t)S(s)$.

Similarly, we can prove that $S(0) = I$ and s-$\lim_{t\to 0^+} S(t)\phi = \phi$. The latter result involves interchanging two limits. This is permissible since the limit in (15) exists uniformly in t. We can also show that for each $t \geq 0$, $S(t)$ is a nonnegative isometry which satisfies the condition in part (ii). (See [9, pp. 48–51] for details.)

(iii) We let $n \to \infty$ in (16) to obtain the result.

(iv) From part (iii), $S(t)\phi = S_m(t)\phi$ for all $\phi \in \mathcal{L}_m$. By Lemma 3.1(ii), we note that $S_m(t) : \mathcal{L}_m \to \mathcal{L}_m$ is invariant under $S_m(t)$, and so $\mathcal{L}_m$ $(m > 0)$ and $\bigcup_m \mathcal{L}_m$ are invariant under $S(t)$. □

5. The infinitesimal generator of the limit semigroup. The infinitesimal generator, $\boldsymbol{\mathcal{A}}$, of the limit semigroup, $\{S(t)\}_{t\geq 0}$, is not necessarily the same as the operator A defined by (8). The limit semigroup gives rise to a solution $u(t) = S(t)f,\ t \geq 0$, of the ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= \boldsymbol{\mathcal{A}}u(t), \quad t > 0, \\ u(0) &= f, \end{aligned} \tag{17}$$

for $\boldsymbol{\mathcal{A}}$, but this solution u need not be a solution of our original ACP (9) for A. In this section, we show that $\boldsymbol{\mathcal{A}}$ and A coincide on a dense domain. We begin by defining two spaces that we shall require in the following analysis.

Definition 5.1. *Let the function a and the operators A and P_n be defined by* (5), (8), *and* (11), *respectively. We define the space $\mathcal{D}_a$ by*

$$\mathcal{D}_a = \left\{\phi \in L_{1,-1}\colon \int_0^\infty xa(x)|\phi(x)|\,dx < \infty\right\} \tag{18}$$

and the space $\mathbf{D}$ *by*

$$\mathbf{D} = \{\phi \in L_{1,-1}\colon \underset{n\to\infty}{\text{s-lim}}\, AP_n\phi \text{ exists in } L_{1,-1}\}. \tag{19}$$

Lemma 5.2.

(i) *For all $\phi \in \mathbf{D}$, $\underset{n\to\infty}{\text{s-lim}}\, AP_n\phi = A\phi$.*

(ii) *Let $\mathcal{L}_m$ be defined by* (10). *Then $\bigcup_m \mathcal{L}_m \subseteq \mathcal{D}_a \subseteq \mathbf{D}$.*

Proof.

(i) Let $\phi \in \mathbf{D}$. Then s-$\lim_{n\to\infty} AP_n\phi$ exists and equals ψ, say. Since $AP_n\phi \to \psi$ in $L_{1,-1}$, there exists a subsequence such that $[AP_{n_k}\phi](x) \to \psi(x)$ a.e. Now

$$[AP_{n_k}\phi](x) = \begin{cases} \int_x^{n_k} \gamma(y,x)\phi(y)dy - a(x)\phi(x), & \text{if } x < n_k, \\ 0, & \text{if } x \geq n_k, \end{cases}$$

and so $\lim_{n_k\to\infty}[AP_{n_k}\phi](x) = [A\phi](x)$. The pointwise limit is unique. Thus $\psi \equiv A\phi$ and s-$\lim_{n\to\infty} AP_n\phi = A\phi$, for all $\phi \in \mathbf{D}$.

(ii) Let $\phi \in \bigcup_m \mathcal{L}_m$. Then $\phi \in \mathcal{L}_m$ for some m and

$$\begin{aligned} \int_0^\infty xa(x)|\phi(x)|dx &\leq \sup_{x\in[0,m]} a(x) \int_0^m x|\phi(x)|dx \\ &< \infty \qquad \text{(since } a \text{ satisfies (H3))}. \end{aligned}$$

Hence $\phi \in \mathcal{D}_a$ and so $\bigcup_m \mathcal{L}_m \subseteq \mathcal{D}_a$.

Now let $\phi \in \mathcal{D}_a$. Using Fubini's theorem, we can show that

$$\|A\phi\|_{1,-1} \leq 2\int_0^\infty xa(x)|\phi(x)|\,dx < \infty,$$

and so

$$\begin{aligned}\|A\phi - AP_n\phi\|_{1,-1} &\leq 2\int_0^\infty xa(x)|\phi(x) - (P_n\phi)(x)|\,dx\\ &\to 0 \qquad \text{as } n \to \infty.\end{aligned}$$

Therefore, for $\phi \in \mathcal{D}_a$, s-$\lim_{n\to\infty} AP_n\phi$ exists and equals $A\phi$. Hence $\mathcal{D}_a \subseteq \mathbf{D}$. □

THEOREM 5.3. *Let $\boldsymbol{\mathcal{A}}$ be the generator of the limit semigroup $\{S(t)\}_{t\geq 0}$ and let A be the operator defined by* (8).

(i) *The operators $\boldsymbol{\mathcal{A}}$ and A are equivalent on the space $\bigcup_m \mathcal{L}_m$, where $\mathcal{L}_m$ is given by* (10). *Moreover, $\bigcup_m \mathcal{L}_m$ is a core for $\boldsymbol{\mathcal{A}}$, i.e.,*

$$\boldsymbol{\mathcal{A}} = \overline{\boldsymbol{\mathcal{A}}|_{\bigcup_m \mathcal{L}_m}}.$$

(ii) *The generator $\boldsymbol{\mathcal{A}}$ and the operator A coincide on $\mathbf{D}$.*

(iii) *The spaces $\mathcal{D}_a$ and $\mathbf{D}$ are also cores for $\boldsymbol{\mathcal{A}}$. So*

$$\boldsymbol{\mathcal{A}} = \overline{\boldsymbol{\mathcal{A}}|_{\bigcup_m \mathcal{L}_m}} = \overline{\boldsymbol{\mathcal{A}}|_{\mathcal{D}_a}} = \overline{\boldsymbol{\mathcal{A}}|_{\mathbf{D}}}.$$

(iv) *The family $\{S(t)\}_{t\geq 0}$ is the unique C_0-semigroup such that $\boldsymbol{\mathcal{A}}\phi = A\phi$ for all $\phi \in \mathbf{D}$, and $S(t)\mathcal{L}_m \subset \mathcal{L}_m$ for all $m > 0$, $t \geq 0$.*

Proof.

(i) From Lemma 4.1(iii), $S(t)\phi = S_m(t)\phi$ for all $\phi \in \mathcal{L}_m$, $m > 0$. By definition, for all $\phi \in \mathcal{L}_m$, $m > 0$,

$$\begin{aligned}\boldsymbol{\mathcal{A}}\phi &= \operatorname*{s-lim}_{t\to 0^+} \frac{S(t)\phi - \phi}{t}\\ &= \operatorname*{s-lim}_{t\to 0^+} \frac{S_m(t)\phi - \phi}{t}\\ &= AP_m\phi \quad \text{(since } AP_m \text{ generates } \{S_m(t)\}_{t\geq 0} \text{ by Lemma 3.1)}\\ &= A\phi.\end{aligned}$$

Hence $\boldsymbol{\mathcal{A}} = A$ on $\bigcup_m \mathcal{L}_m$.

Now, $\bigcup_m \mathcal{L}_m$ is a dense subspace of $L_{1,-1}$, $\bigcup_m \mathcal{L}_m \subseteq \mathcal{D}(\boldsymbol{\mathcal{A}})$, and (by Lemma 4.1(iv)) $S(t) : \bigcup_m \mathcal{L}_m \to \bigcup_m \mathcal{L}_m$. Applying the core theorem [17, Thm. X.49, pp. 241–242], we obtain the required result.

(ii) Let $\phi \in \mathbf{D}$. Then s-$\lim_{n\to\infty} AP_n\phi$ exists and equals $A\phi$ (by Lemma 5.2(i)). Since $P_n\phi \in \mathcal{L}_n$, by part (i), we note that

$$\operatorname*{s-lim}_{n\to\infty} \boldsymbol{\mathcal{A}}P_n\phi = \operatorname*{s-lim}_{n\to\infty} AP_n\phi = A\phi.$$

Therefore we have that $P_n\phi \to \phi$ and $\boldsymbol{\mathcal{A}}P_n\phi \to A\phi$, as $n \to \infty$. Since the generator $\boldsymbol{\mathcal{A}}$ is closed, we deduce that $\mathbf{D} \subseteq \mathcal{D}(\boldsymbol{\mathcal{A}})$ and $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathbf{D}$.

(iii) Let $G(\boldsymbol{\mathcal{A}})$ denote the graph of $\boldsymbol{\mathcal{A}}$. Then

$$\overline{G(\boldsymbol{\mathcal{A}}|_{\mathbf{D}})} \subseteq \overline{G(\boldsymbol{\mathcal{A}})} = G(\boldsymbol{\mathcal{A}}),$$

since $\mathbf{D} \subseteq \mathcal{D}(\boldsymbol{\mathcal{A}})$. However, because $\mathbf{D} \supseteq \bigcup_m \mathcal{L}_m$, we obtain

$$\overline{G(\boldsymbol{\mathcal{A}}|_{\mathbf{D}})} \supseteq \overline{G(\boldsymbol{\mathcal{A}}|_{\bigcup_m \mathcal{L}_m})} = G(\boldsymbol{\mathcal{A}})$$

by part (i). Thus $G(\boldsymbol{\mathcal{A}}) = \overline{G(\boldsymbol{\mathcal{A}}|_{\mathbf{D}})}$. By the definition of closure, $\overline{G(\boldsymbol{\mathcal{A}}|_{\mathbf{D}})} = G(\overline{\boldsymbol{\mathcal{A}}|_{\mathbf{D}}})$, and so $\boldsymbol{\mathcal{A}} = \overline{\boldsymbol{\mathcal{A}}|_{\mathbf{D}}}$. Hence $\mathbf{D}$ is a core for $\boldsymbol{\mathcal{A}}$. A similar argument holds for $\mathcal{D}_a$, since $\bigcup_m \mathcal{L}_m \subseteq \mathcal{D}_a \subseteq \mathbf{D} \subseteq \mathcal{D}(\boldsymbol{\mathcal{A}})$.

(iv) By [15, Thm. 2.6, p. 6], $\{S(t)\}_{t\geq 0}$ is the unique semigroup generated by $\boldsymbol{\mathcal{A}}$. Suppose there exists another generator $\mathcal{B}$, which gives rise to the C$_0$-semigroup $\{T(t)\}_{t\geq 0}$, such that $\mathcal{B} \neq \boldsymbol{\mathcal{A}}$, $\mathcal{B} = A$ on $\mathbf{D}$, and $T(t)\mathcal{L}_m \subseteq \mathcal{L}_m$, for all $m > 0$, $t \geq 0$. By the core theorem applied to $\mathcal{B}$ and $\bigcup_m \mathcal{L}_m$, we obtain

$$\mathcal{B} = \overline{\mathcal{B}|_{\bigcup_m \mathcal{L}_m}}.$$

Since $\bigcup_m \mathcal{L}_m \subseteq \mathbf{D}$, by the above assumptions on $\mathcal{B}$, we note that $\mathcal{B}\phi = A\phi = \boldsymbol{\mathcal{A}}\phi$ for all $\phi \in \bigcup_m \mathcal{L}_m$. So

$$\mathcal{B} = \overline{\boldsymbol{\mathcal{A}}|_{\bigcup_m \mathcal{L}_m}} = \boldsymbol{\mathcal{A}}.$$

This is a contradiction. Thus $\{S(t)\}_{t\geq 0}$ is the only semigroup that leaves $\mathcal{L}_m$ $(m > 0)$ invariant and whose generator $\boldsymbol{\mathcal{A}}$ coincides with the operator A on $\mathbf{D}$. □

We have shown that the generator $\boldsymbol{\mathcal{A}}$ and the operator A, defined by (8), coincide on the dense space $\mathbf{D}$. Ideally we would like $\boldsymbol{\mathcal{A}}$ and A to coincide on $\mathcal{D}(\boldsymbol{\mathcal{A}})$ so that if we considered A to be defined on $\mathcal{D}(\boldsymbol{\mathcal{A}})$ instead of on its maximal domain the ACPs (9) and (17) would become the same problem and for suitable initial data, $u(t) = S(t)f$, $t \geq 0$, would be the solution. However, so far, we have been unable to determine the precise nature of $\mathcal{D}(\boldsymbol{\mathcal{A}})$ and whether $\mathcal{D}(\boldsymbol{\mathcal{A}}) \subseteq D_{\max}(A)$ without imposing extra conditions on the fragmentation kernel $\gamma(x, y)$, as in the next theorem.

Theorem 5.4. *Let X be a Banach space of functions on $(0, \infty)$ with the property that strong convergence in X implies pointwise convergence almost everywhere. In addition to the hypotheses* (H1)–(H3), *let the fragmentation kernel $\gamma(x, y)$ be such that the operator A defined by* (8) *is a bounded operator from $L_{1,-1}$ into X. Then the operator A coincides with the infinitesimal generator, $\boldsymbol{\mathcal{A}}$, of the limit semigroup on the domain, $\mathcal{D}(\boldsymbol{\mathcal{A}})$, of the generator. Hence $\boldsymbol{\mathcal{A}}$ is a restriction of A.*

Proof. Let $\phi \in \mathcal{D}(\boldsymbol{\mathcal{A}})$. Since $\mathbf{D}$ is a core for $\boldsymbol{\mathcal{A}}$, there exists a sequence $\{\phi_n\} \subseteq \mathbf{D}$ such that $\|\phi_n - \phi\|_{1,-1} \to 0$ and $\|\boldsymbol{\mathcal{A}}\phi_n - \boldsymbol{\mathcal{A}}\phi\|_{1,-1} \to 0$ as $n \to \infty$. Hence there exists a subsequence $\{\phi_{n_k}\}$ such that $[\boldsymbol{\mathcal{A}}\phi_{n_k}](x) = [A\phi_{n_k}](x) \to [\boldsymbol{\mathcal{A}}\phi](x)$ for a.e. $x > 0$, as $n_k \to \infty$, where we have used the fact that $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathbf{D}$.

Since $A \in B(L_{1,-1}, X)$,

$$A\phi = \operatorname*{s-lim}_{n_k\to\infty} A\phi_{n_k},$$

where the strong limit is with respect to the norm on X. By the assumptions on X, there exists a subsequence of the subsequence (which we also denote by ϕ_{n_k}) such that $[A\phi_{n_k}](x) \to [A\phi](x)$ for a.e. $x > 0$, as $n_k \to \infty$. Since the pointwise limit is unique, $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathcal{D}(\boldsymbol{\mathcal{A}})$, and hence $\boldsymbol{\mathcal{A}} = A|_{\mathcal{D}(\boldsymbol{\mathcal{A}})}$. □

Sufficient conditions on γ for the above theorem to be applicable are now given.

Example 5.5. *Using the generalized Young's inequality* [5, pp. 13–14]*, it is not difficult to show that for fragmentation kernels γ satisfying*

$$\int_x^\infty \frac{1}{y}\gamma(y, x)\, dy \leq C, \qquad \textit{for all } x > 0,$$

and

$$\int_0^y \frac{1}{y}\gamma(y, x)\, dx \leq C, \qquad \textit{for all } y > 0,$$

for some constant $C > 0$, the operator $A \in B(L_{1,-1}, L_1)$. An example of such a γ is $\gamma(y, x) = x/y$. Note that the above conditions imply that (H3) *is valid.*

THEOREM 5.6. *Let the fragmentation kernel γ satisfy* (H1) *and* (H2) *and be such that*

$$\gamma(x, y) \leq C x^{\beta-1} \Psi\left(\frac{y}{x}\right), \tag{20}$$

where $\beta > 0$ and the function Ψ is such that

$$\int_0^1 t^{1-\beta} \Psi(t)\, dt = C' < \infty. \tag{21}$$

Then $A \in B(L_{1,-1}, L_{1,\beta-1})$, where $L_{1,\beta-1}$ is defined as in [7, p. 2].

Proof. From (8) and (5), the operator A can be expressed as the sum $B_1 + B_2$, where for $\phi \in D_{\max}(A)$ and $x > 0$,

$$(B_1\phi)(x) = \int_x^\infty \gamma(y, x)\phi(y)\, dy \quad \text{and} \quad (B_2\phi)(x) = -a(x)\phi(x).$$

We consider $\|B_1\phi\|_{1,-1}$ and $\|B_2\phi\|_{1,-1}$ separately. The result follows from the estimates (20) and (21) on making a suitable substitution or changing the order of integration, as appropriate. □

EXAMPLE 5.7. *The following examples are special cases of the above theorem.*

(i) *Let $\Psi(t) = t^\lambda$ and $\beta = 2\lambda + 1$ so that $\gamma(x, y) \leq C(xy)^\lambda$. If $-\frac{1}{2} < \lambda < 1$ then $A \in B(L_{1,-1}, L_{1,2\lambda})$. In particular, choosing $\lambda = 0$, we see that $A \in B(L_{1,-1}, L_{1,0})$ whenever $\gamma(x, y) \leq C$ for some constant $C < \infty$. Also, when $\gamma(x, y) = (xy)^\lambda$, the operator B_1 (defined above) reduces to $x^{2\lambda+1} K_1^{-\lambda-1,1}\phi$ where $K_1^{-\lambda-1,1}$ is an Erdélyi–Kober operator; see* [7, p. 39].

(ii) *Let $\Psi(t) = (1 - t)^{\alpha-1}$ $(0 < \alpha < 2)$ and $\beta = \alpha$. Then $\gamma(x, y) \leq C(x - y)^{\alpha-1}$ and $A \in B(L_{1,-1}, L_{1,\alpha-1})$. If $\gamma(x, y) = \frac{1}{\Gamma(\alpha)}(x - y)^{\alpha-1}$, then the operator B_1 is the Weyl fractional integral K_1^α of order α; see* [7, p. 36].

6. Existence and uniqueness results. We investigate the existence and uniqueness of solutions to the ACP (17) for $\boldsymbol{\mathcal{A}}$ and the ACP (9) for A. We also consider the ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= \boldsymbol{\mathcal{A}}u(t) + q(t), \quad t > 0, \\ u(0) &= f, \end{aligned} \tag{22}$$

where q is a vector-valued function. The inhomogeneous term $q(t)$ can be interpreted as arising from a source term $q(x, t)$ being introduced on the right-hand side of the pure fragmentation equation (2).

THEOREM 6.1. *Let $f \in \mathcal{D}(\boldsymbol{\mathcal{A}})$, $f \geq 0$. Then the ACP* (17) *has a unique, strongly differentiable, nonnegative, mass-conserving solution given by $u(t) = S(t)f$ for all $t \geq 0$.*

Proof. The result follows by [15, Thm. 1.3, pp. 102–103] and Lemma 4.1(i), (ii) since, by definition, $\boldsymbol{\mathcal{A}}$ generates $\{S(t)\}_{t\geq 0}$. □

COROLLARY 6.2. *Suppose that one or both of the following conditions is valid:*

(i) *the generator $\boldsymbol{\mathcal{A}}$ coincides with A on $\mathcal{D}(\boldsymbol{\mathcal{A}})$;*

(ii) *$f \in \mathbf{D}$ and is such that $S(t)f \in \mathbf{D}$ for all $t \geq 0$.*

Then Theorem 6.1 *is also valid for the ACP* (9).

Proof. This is an immediate consequence of the previous theorem since, respectively,

(i) $u(t) \in \mathcal{D}(\boldsymbol{\mathcal{A}})$ for all time and $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathcal{D}(\boldsymbol{\mathcal{A}})$;

(ii) $u(t) \in \mathbf{D}$ for all time and $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathbf{D}$ (by Theorem 5.3(ii)). □

Note. For arbitrary initial data $f \in \mathbf{D}$, there is no guarantee that $S(t)f \in \mathbf{D}$ for all time. Consequently, for such data, we can only assert that the ACP (9) has at most one strongly differentiable, nonnegative, mass-conserving solution given by $u(t) = S(t)f,\ t \geq 0$.

COROLLARY 6.3. *Let $f \in \bigcup_m \mathcal{L}_m,\ f \geq 0$. Then there exists a unique strongly differentiable, nonnegative, mass-conserving solution to the ACP* (9) *such that $u(t) \in \bigcup_m \mathcal{L}_m$ for all $t \geq 0$. This strong solution is given by $u(t) = S(t)f,\ t \geq 0$.*

Proof. By Lemma 4.1(iv), $\bigcup_m \mathcal{L}_m$ is invariant under $S(t)$. The result follows from Corollary 6.2 since $\bigcup_m \mathcal{L}_m \subseteq \mathbf{D}$. □

The above analysis established the existence and uniqueness of a C_0-semigroup $\{S(t)\}_{t\geq 0}$ with infinitesimal generator $\boldsymbol{\mathcal{A}}$ that coincides with the operator A on a dense set $\mathbf{D}$, such that $\mathcal{L}_m$ is invariant under $S(t),\ t \geq 0$. This C_0-semigroup gives rise to a strong solution of the ACP (17) for $\boldsymbol{\mathcal{A}}$. However, in general, we do not have a precise form for $\boldsymbol{\mathcal{A}}$, and so it is difficult to interpret (meaningfully) the abstract problem we are actually solving. In certain circumstances (see Theorem 5.4) $\boldsymbol{\mathcal{A}}$ is a restriction of the operator A defined by (8), and so, by taking A to be defined on the set $\mathcal{D}(\boldsymbol{\mathcal{A}})$ instead of its maximal domain, results relating to the ACP (17) are also valid for the ACP (9).

In the remainder of this section, we state results for the ACPs (17) and (22), involving the "abstract" operator $\boldsymbol{\mathcal{A}}$. We note that in certain cases these results apply to the operator A and hence to the abstract formulation of the pure fragmentation equation (2) with a source term $q(x,t)$.

DEFINITION 6.4. *Fix $T < \infty$. A function $u \in C([0,T],X)$ is a weak solution of* (22) *on $[0,T]$ if for every $v \in \mathcal{D}(\boldsymbol{\mathcal{A}}^*)$ the function $\langle u(t), v\rangle$ is absolutely continuous on $[0,T]$ and*

$$\frac{d}{dt}\langle u(t), v\rangle = \langle u(t), \boldsymbol{\mathcal{A}}^* v\rangle + \langle q(t), v\rangle, \quad \textit{for almost all } t \in [0,T],$$

and $u(0) = f$.

THEOREM 6.5 (existence and uniqueness of a weak solution). *Let the initial data $f \in L_{1,-1},\ f \geq 0$.*

(i) *The ACP* (17) *has a unique weak solution u satisfying $u(0) = f$. The weak solution is given by $u(t) = S(t)f,\ t \geq 0$, is nonnegative, and conserves mass.*

(ii) *For each fixed $T > 0$, if $q \in L_1([0,T), L_{1,-1})$, then the ACP* (22) *has a unique weak solution u on $[0,T]$ satisfying $u(0) = f$. This weak solution is given by*

$$u(t) = S(t)f + \int_0^t S(t-s)q(s)\,ds, \quad t \in [0,T].$$

If $q \geq 0$, then the weak solution to (22) *is nonnegative.*

Proof. The existence and uniqueness results follow immediately from [3], since $\boldsymbol{\mathcal{A}}$ generates the C_0-semigroup $\{S(t)\}_{t\geq 0}$. The weak solution of the ACP (17) is nonnegative and mass conserving because the operators $S(t),\ t \geq 0$, are nonnegative isometries. Likewise, the weak solution of (22) on $[0,T]$ is nonnegative whenever $q \geq 0$. □

7. Existence and uniqueness of solutions to the pure fragmentation equation. We consider the implications on the original rate equation (2) of having existence and uniqueness of solutions to the ACP (17). We also compare our definition of a solution with that adopted in [18, 19]. This enables us to make comparisons between the existence and uniqueness results obtained here and those obtained in [18, 19] under a different set of hypotheses.

THEOREM 7.1. *Let γ satisfy* (H1)–(H3) *and let $f \in \bigcup_m \mathcal{L}_m$, $f \geq 0$. Then there exists a function u, measurable on the product set $(0,\infty) \times [0,\infty)$, such that*

(i) *$u(x,t) \geq 0$ for all $t \geq 0$ and a.e. $x > 0$;*

(ii) *$u(x,0) = f(x)$ for a.e. $x > 0$;*

(iii) *for each fixed $t \geq 0$, $\int_0^\infty xu(x,t)\,dx < \infty$;*

(iv) *$\int_0^\infty xu(x,t)\,dx = \int_0^\infty xf(x)\,dx$ for all $t \geq 0$;*

(v) *$u(x,t)$ satisfies the pure fragmentation equation* (2) *almost everywhere.*

Proof. By Corollary 6.3, the ACP (17) has a strongly continuously differentiable solution $u(t)$, $t \geq 0$. By [6, Thm. 3.4.2, pp. 70–71], since $L_{1,-1}$ is of type L, there exists a function u measurable on the product set $(0,\infty) \times [0,\infty)$ such that

(a) $u(x,t)$ is absolutely continuous for each $x \in (0,\infty)$ and $u(x,t) = [u(t)](x)$ for each $t \geq 0$;

(b) $\frac{\partial}{\partial t}u(x,t)$ exists a.e. in $(0,\infty)\times[0,\infty)$ and $\frac{\partial}{\partial t}u(x,t) = \left[\frac{d}{dt}u(t)\right](x)$ for all $t \geq 0$. Now,

$$\left[\frac{d}{dt}u(t)\right](x) = [\boldsymbol{\mathcal{A}}u(t)](x) = [Au(t)](x)$$

since $u(t) \in \bigcup_m \mathcal{L}_m$ (Corollary 6.3) and $\boldsymbol{\mathcal{A}} \equiv A$ on $\mathbf{D} \supseteq \bigcup_m \mathcal{L}_m$ (Theorem 5.3(ii)). By definition, for $t > 0$,

$$\begin{aligned}[Au(t)](x) &= \int_x^\infty \gamma(y,x)[u(t)](y)dy - [u(t)](x)\int_0^x \frac{y}{x}\gamma(x,y)\,dy\\ &= \int_x^\infty \gamma(y,x)u(y,t)dy - u(x,t)\int_0^x \frac{y}{x}\gamma(x,y)\,dy.\end{aligned}$$

Thus

$$\frac{\partial}{\partial t}u(x,t) = \int_x^\infty \gamma(y,x)u(y,t)dy - u(x,t)\int_0^x \frac{y}{x}\gamma(x,y)\,dy, \quad \text{a.e. } x > 0,\ t > 0;$$

i.e., $u(x,t)$ satisfies (2) almost everywhere.

On setting $u(x,t) = [u(t)](x)$, for a.e. $x > 0$, $t \geq 0$, parts (i)–(iv) follow immediately from the corresponding properties of $u(t)$. □

COROLLARY 7.2. *Let $f \in \bigcup_m \mathcal{L}_m$, $f \geq 0$, and let u be a solution of the pure fragmentation equation* (2) *in the sense described in Theorem* 7.1. *Then this is the only such solution which is also a solution in the strong sense, i.e., such that the vector-valued function u defined by $u(t) = u(.,t)$, $t \geq 0$, is a strong solution of the ACP* (17).

Proof. This is a direct consequence of the uniqueness of the strong solution for $f \in \bigcup_m \mathcal{L}_m$ (Corollary 6.3). □

The additional requirement that a solution of the pure fragmentation equation (2) should also satisfy the equation in a strong sense provides one means of uniquely identifying a solution to (2) for initial data $f \in \bigcup_m \mathcal{L}_m$. For arbitrary initial data $f \in L_{1,-1}$, $f \geq 0$, we have the existence and uniqueness of a weak solution to the ACP (17) (but not so, in general, for the ACP (9)) given by $u(t) = S(t)f$, $t \geq 0$. Note that for all $t \geq 0$,

$$S(t)f = \operatorname*{s-lim}_{n\to\infty} S(t)P_n f.$$

Therefore, the solution u is a limit of solutions $\{u_n\}$, given by

$$u_n(t) = S(t)P_n f, \qquad t \geq 0,$$

which satisfy the ACP (17) strongly (since $P_n f \in \bigcup_m \mathcal{L}_m \subseteq \mathcal{D}(\boldsymbol{\mathcal{A}})$). We now show that for arbitrary initial data $f \in L_{1,-1}, f \geq 0$, the function $u(x,t) = [u(t)](x),\ x > 0,\ t \geq 0$, satisfies an integral version of the pure fragmentation equation (2).

THEOREM 7.3. *Let the initial data $f \in L_{1,-1},\ f \geq 0$. Then the function $u\colon (0,\infty) \times [0,\infty) \to \mathbf{R}$ defined by $u(x,t) = [S(t)f](x),\ x > 0,\ t \geq 0$ satisfies*

$$u(x,t) = f(x) + \int_0^t \int_x^\infty \gamma(y,x) u(y,\sigma)\, dy\, d\sigma - \int_0^t a(x) u(x,\sigma)\, d\sigma \tag{23}$$

for all $t \geq 0$ and a.e. $x > 0$. Moreover, u is nonnegative and

$$\int_0^\infty x u(x,t)\, dx = \int_0^\infty x f(x)\, dx \quad \textit{for all } t \geq 0.$$

Proof. Let $u_n(t) = S(t)P_n f$ and $u(t) = S(t)f$. By [6, pp. 69–71], for each fixed $t \geq 0$, we can write $[u_{n_k}(t)](x) = u_{n_k}(x,t)$ and $[u(t)](x) = u(x,t)$, where the scalar-valued functions u_{n_k} and u appearing on the right-hand sides are measurable on the product set $(0,\infty) \times [0,\infty)$. Since $u_n(t) \to u(t)$ strongly in $L_{1,-1}$ uniformly with respect to t, for each fixed $t \geq 0$ there exists a subsequence $\{n_k\}$ independent of t such that

$$u_{n_k}(x,t) \to u(x,t) \qquad \text{as } n_k \to \infty \quad \text{for a.e. } x > 0,$$

where the set of measure zero may depend on t. However, the uniform convergence of the sequence $\{u_{n_k}(t)\}$ in $L_{1,-1}$ also means that

$$\int_0^T \int_0^\infty x |u_{n_k}(x,t) - u(x,t)|\, dx\, dt \to 0 \qquad \text{as } n_k \to \infty$$

for each finite $T > 0$, and from this we deduce that there exists a subsequence of the subsequence, which we also denote by $\{u_{n_k}(t)\}$, such that

$$u_{n_k}(x,t) \to u(x,t) \qquad \text{as } n_k \to \infty \quad \text{for a.e. } (x,t) \in (0,\infty) \times [0,T].$$

By Theorem 7.1, because $P_{n_k} f \in \mathcal{L}_{n_k} \subset \bigcup_m \mathcal{L}_m$, we deduce that $u_{n_k}(x,t)$ satisfies

$$u_{n_k}(x,t) = (P_{n_k} f)(x) + \int_0^t \int_x^\infty \gamma(y,x) u_{n_k}(y,\sigma)\, dy\, d\sigma - a(x) \int_0^t u_{n_k}(x,\sigma)\, d\sigma \tag{24}$$

for each fixed $t \geq 0$ and a.e. $x > 0$. Note that $u_{n_k}(x,t) = 0$ for $x \geq n_k$, since $\mathcal{L}_{n_k}$ is invariant under $S(t)$ (by Lemma 4.1(iv)).

Taking limits as $n_k \to \infty$ on both sides of (24), we get

$$u(x,t) = f(x) + \lim_{n_k \to \infty} \left\{ \int_0^t \int_x^\infty \gamma(y,x) u_{n_k}(y,\sigma) dy d\sigma - a(x) \int_0^t u_{n_k}(x,\sigma) d\sigma \right\} \tag{25}$$

for each fixed $t \geq 0$ and a.e. $x > 0$, where we know that the limit on the right exists because the limit on the left exists.

Now fix $t \geq 0$. For almost all fixed $x > 0$, we can show that the sequence $\{a(x) u_{n_k}(x, \,.\,)\}$ is monotonic increasing, is bounded above by the $L_1[0,t]$-function $a(x)u(x, \,.\,)$, and converges pointwise a.e. on $[0,t]$ to $a(x)u(x,\sigma)$. By the Lebesgue dominated convergence theorem, for each fixed $t \geq 0$ and almost all fixed $x > 0$,

$$\lim_{n_k \to \infty} \int_0^t a(x) u_{n_k}(x,\sigma)\, d\sigma = \int_0^t a(x) u(x,\sigma)\, d\sigma.$$

Returning to (25), we know now that $\lim_{n_k\to\infty}\int_0^t\int_x^\infty \gamma(y,x)u_{n_k}(y,\sigma)\,dy\,d\sigma$ exists for each fixed $t \geq 0$ and almost all fixed $x > 0$. For fixed $y \in (x,\infty)$ and $\sigma \in [0,t]$, the sequence $\{\gamma(y,x)u_{n_k}(y,\sigma)\}$ is monotonic and converges a.e. on $(x,\infty)\times[0,t]$ to $\gamma(y,x)u(y,\sigma)$. Hence, by the Fubini–Tonelli–Hobson theorem and the Levi theorem for sequences of Lebesgue integrable functions [2, Thm. 10.24, pp. 267–268 and p. 407],

$$\lim_{n_k\to\infty}\int_0^t\int_x^\infty \gamma(y,x)u_{n_k}(y,\sigma)\,dy\,d\sigma = \int_0^t\int_x^\infty \gamma(y,x)u(y,\sigma)\,dy\,d\sigma.$$

We have thus shown that u satisfies (19). The nonnegativity and mass conservation follow from Lemma 4.1(i)–(ii). □

COROLLARY 7.4. *There is only one solution u of the integral version of the rate equation* (19) *such that the vector-valued function $u(t)$ defined by $u(t) = u(.,t)$, $t \geq 0$, is a weak solution of the ACP* (17). *This solution is given by*

$$u(x,t) = [S(t)f](x), \quad \textit{for a.e. } x > 0,\ t \geq 0.$$

Proof. This follows since the weak solution is unique (Theorem 6.5). □

COROLLARY 7.5. *The solution u of the integral version* (23) *satisfies the pure fragmentation equation* (2) *for a.e. $t > 0$, $x > 0$ and satisfies the initial condition $u(x,0) = f(x)$.*

Proof. Since $u(x,t)$ satisfies (23) for all $t \geq 0$ and a.e. $x > 0$, $u(x,.)$ is absolutely continuous on $[0,\infty)$ and so $u(x,t)$ satisfies (2) for a.e. $t > 0$, $x > 0$. □

Remark 7.6.

1. We are able to obtain existence and uniqueness results for the pure fragmentation equation (2) regardless of whether or not the abstract problems for $\boldsymbol{\mathcal{A}}$ and A coincide.

2. In [1], although use is made of the pointwise solution $u(x,t) = [u(t)](x)$, $x > 0$, $t \geq 0$, no justification is provided. The above discussion makes this step rigorous.

3. Comparisons can now be made with Stewart's existence and uniqueness results [18, 19] (see section 1) for the case of pure fragmentation ($K \equiv 0$). Under the hypotheses (H1)–(H3) on the fragmentation kernel, we have established the existence of a solution $u(x,t) = [S(t)f](x)$, $x > 0$, $t \geq 0$, which is also a solution in the sense described in [18]. By restricting our solution of (2) to satisfy (9) in a weak sense, we obtain uniqueness without imposing further conditions on the kernels (compare with [19]). Thus, in the pure fragmentation case, our results extend those in [18, 19]. However, it should be noted that unbounded fragmentation and unbounded coagulation are dealt with simultaneously in [18], and so it is possible that the same approach applied separately to the pure fragmentation equation would lead to less restrictive conditions on the fragmentation kernel.

REFERENCES

[1] M. AIZENMAN AND T.A. BAK, *Convergence to equilibrium in a system of reacting polymers*, Comm. Math. Phys., 65 (1979), pp. 203–230.

[2] T. M. APOSTOL, *Mathematical Analysis*, in World Student Series (Second edition), Addison–Wesley, Reading, MA, 1974.

[3] J. M. BALL, *Strongly continuous semigroups, weak solutions, and a variation of constants formula*, Proc. Amer. Math. Soc., 63 (1977), pp. 370–373.

[4] A. F. FILIPPOV, *On the distribution of the sizes of particles which undergo splitting*, Theory Probab. Appl., 6 (1961), pp. 275–294.

[5] G. B. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press and University of Tokyo Press, Princeton, NJ, 1976.

[6] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, American Mathematical Society Colloquium Publications, AMS, Providence, RI, 1957.

[7] A. C. McBride, *Fractional Calculus and Integral Transforms of Generalised Functions*, Research Notes in Mathematics, Pitman, London, 1979.

[8] A. C. McBride, D. J. McLaughlin, and W. Lamb, *On a coagulation-fragmentation equation*, in Lecture Notes in Pure and Applied Mathematics: Evolution Equations, Vol. 168, G. Ferreyra, G.R. Goldstein, and F. Neubrander, eds., Marcel Dekker, New York, 1995, pp. 277–286.

[9] D. J. McLaughlin *Coagulation and Fragmentation Models: A Semigroup Approach*, Ph.D. thesis, Univ. of Strathclyde, Glasgow, Scotland, February 1995.

[10] D. J. McLaughlin, W. Lamb, and A. C. McBride, *A Coagulation-Fragmentation Equation: Pure Fragmentation*, Res. Rep. No. 16, Dept. of Mathematics, Univ. of Strathclyde, Glasgow, Scotland, 1993.

[11] D. J. McLaughlin, W. Lamb, and A. C. McBride, *A Coagulation-Fragmentation Equation: The Full Equation*, Res. Rep. No. 17, Dept. of Mathematics, Univ. of Strathclyde, Glasgow, Scotland, 1993.

[12] D. J. McLaughlin, W. Lamb, and A. C. McBride, *A Time-Dependent Fragmentation Equation*, Res. Rep. No. 18, Dept. of Mathematics, Univ. of Strathclyde, Glasgow, Scotland, 1993.

[13] Z. A. Melzak, *A scalar transport equation*, Trans. Amer. Math. Soc., 85 (1957), pp. 547–560.

[14] Z. A. Melzak, *A scalar transport equation* II, Mich. Math. J., 4 (1957), pp. 193–206.

[15] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

[16] J. D. Pryce, *Basic Methods of Linear Functional Analysis*, Hutchinson, London, 1973.

[17] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, London, 1975.

[18] I.W. Stewart, *A global existence theorem for the general coagulation-fragmentation equation with unbounded kernels*, Math. Meth. Appl. Sci., 11 (1989), pp. 627–648.

[19] I.W. Stewart, *A uniqueness theorem for the coagulation-fragmentation equation*, Math. Proc. Camb. Philos. Soc., 107 (1990), pp. 573–578.

[20] R.M. Ziff and E.D. McGrady, *The kinetics of cluster fragmentation and depolymerisation*, J. Phys. A Math. Gen., 18 (1985), pp. 3027–3037.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1173–1190, September 1997

AN EXISTENCE AND UNIQUENESS RESULT FOR A COAGULATION AND MULTIPLE-FRAGMENTATION EQUATION*

D. J. McLAUGHLIN†, W. LAMB†, AND A. C. McBRIDE†

Abstract. Prior knowledge regarding the existence and uniqueness of nonnegative, mass-conserving solutions to a multiple-fragmentation equation is utilized to study a combined coagulation and fragmentation model. The coagulation and fragmentation equation is first recast as an abstract integral equation involving the solution operator associated with the fragmentation part. A contraction mapping argument is then used to prove the existence and uniqueness of a local solution. Detailed investigation of the related iteration scheme yields nonnegativity and mass conservation. The solution is shown to be global.

Key words. semigroup theory, coagulation, multiple fragmentation, contraction mapping

AMS subject classifications. 35F25, 45K05, 45N05, 47D06, 47N20, 80A30

PII. S0036141095291713

1. Introduction.

In this paper, we continue our investigation of the continuous coagulation and multiple-fragmentation equation,

$$\begin{aligned}\frac{\partial}{\partial t}u(x,t) &= \frac{1}{2}\int_0^x K(x-y,y)u(x-y,t)u(y,t)\,dy - u(x,t)\int_0^\infty K(x,y)u(y,t)\,dy\\ &\quad + \int_x^\infty \gamma(y,x)u(y,t)\,dy - u(x,t)\int_0^x \frac{y}{x}\gamma(x,y)\,dy, \quad \text{a.e. } x>0,\ t>0, \qquad (1)\end{aligned}$$

for time-independent kernels γ and K. Results obtained in our earlier paper [5] for the pure fragmentation equation ($K \equiv 0$) are now extended to models where coagulation is also taken into account. Future papers will deal with time-dependent kernels.

We recall that the fragmentation kernel $\gamma(x,y)$ represents the rate at which particles of mass y are produced from fragmenting particles of mass x. This formulation of the fragmentation kernel allows for particles splitting into more than two pieces, yet incorporates the special case of binary fragmentation. The latter case can be described by a symmetric fragmentation kernel $F(x,y)$, obtained from γ by setting $F(x,y) = \gamma(x+y,y)$ for $x, y \geq 0$.

The coagulation kernel $K(x,y)$ is the rate at which particles of mass x coalesce with particles of mass y. We introduce this kernel by assuming that the average number of coalescences between particles having mass in $(x, x+dx)$ and those having mass in $(y, y+dy)$ is $K(x,y)u(x,t)u(y,t)\,dx\,dy\,dt$ during the time interval $(t, t+dt)$. We assume this kernel is symmetric since the rate at which a particle of mass x coalesces with a particle of mass y is the same as that for a particle of mass y coalescing with one of mass x.

The first term on the right-hand side of equation (1) represents the increase in the number of particles of mass x as the result of particles of mass $x-y$ and mass y ($y \leq x$) merging to form a particle of mass x. The factor 1/2 takes into account

*Received by the editors September 8, 1995; accepted for publication (in revised form) May 1, 1996. The research of the first author was supported by the Engineering and Physical Sciences Research Council of the United Kingdom.

http://www.siam.org/journals/sima/28-5/29171.html

†Department of Mathematics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK (d.j.mclaughlin@strath.ac.uk, w.lamb@strath.ac.uk, a.c.mcbride@strath.ac.uk).

that either a particle of mass $x-y$ coalesces with one of mass y or vice versa. The second term accounts for the loss of particles of mass x because they have coalesced with particles of mass y, $y \geq 0$. The remaining terms are attributed to fragmentation and arise as described in [5].

Again, the first step is to recast the rate equation, subject to the initial mass distribution $u(x,0) = f(x)$, as an abstract Cauchy problem (ACP). Arguing as in [5], we arrive at the ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= Au(t) + Nu(t), \quad t > 0, \\ u(0) &= f, \end{aligned} \tag{2}$$

where the linear fragmentation operator A and the nonlinear coagulation operator N are defined on suitable domains by

$$[A\phi](x) = \int_x^\infty \gamma(y,x)\phi(y)\,dy - \phi(x)\int_0^x \frac{y}{x}\gamma(x,y)\,dy \tag{3}$$

and

$$[N\phi](x) = \frac{1}{2}\int_0^x K(x-y,y)\phi(x-y)\phi(y)\,dy - \phi(x)\int_0^\infty K(x,y)\phi(y)\,dy, \tag{4}$$

respectively. By treating the initial-value problem (2) as a nonlinear perturbation of the linear ACP

$$\begin{aligned} \frac{d}{dt}u(t) &= Au(t), \quad t > 0, \\ u(0) &= f, \end{aligned} \tag{5}$$

which corresponds to the pure fragmentation equation ($K \equiv 0$), we are able to make use of the results obtained in [5].

We recall from [5] that, for suitable $f \in L_{1,-1}$, the ACP (5) has a strong solution

$$u(t) = S(t)f, \qquad t \geq 0,$$

in the weighted L_1-space

$$L_{1,-1} = \left\{ \phi : \|\phi\|_{1,-1} = \int_0^\infty x|\phi(x)|\,dx < \infty \right\}.$$

The solution operators

$$\{S(t)\}_{t\geq 0} \tag{6}$$

form a nonnegative, C$_0$-semigroup of isometries on $L_{1,-1}$. Moreover, if f is any function in $L_{1,-1}$ with $f \geq 0$, then

$$u(x,t) = [S(t)f](x) \qquad \text{for all } t \geq 0, \text{ a.e. } x > 0, \tag{7}$$

is a nonnegative solution of the integral version of the fragmentation equation,

$$u(x,t) = f(x) + \int_0^t \int_x^\infty \gamma(y,x)u(y,\sigma)\,dy\,d\sigma - \int_0^t \int_0^x \frac{y}{x}\gamma(x,y)\,dy\,u(x,\sigma)\,d\sigma,$$

and satisfies

$$\int_0^\infty xu(x,t)\,dx = \int_0^\infty xf(x)\,dx \qquad \text{for all } t \geq 0. \tag{8}$$

These results are valid provided that the fragmentation kernel γ satisfies the following hypotheses:

(H1) $\gamma(x,y)$ is a nonnegative function on $(0,\infty)\times(0,\infty)$;

(H2) $\gamma(x,y) = 0$ whenever $y > x$;

(H3) the function $a : (0,\infty) \to [0,\infty)$ defined by

$$a(x) = \int_0^x \frac{y}{x}\gamma(x,y)\,dy, \qquad x > 0, \tag{9}$$

is such that $a(x) \leq C_n$ for all $x \in (0,n],\ n > 0$.

Connections between the operator A, defined by (3), and the infinitesimal generator, $\boldsymbol{\mathcal{A}}$, of the semigroup $\{S(t)\}_{t\geq 0}$ on $L_{1,-1}$ are also given in [5]. For example, A and $\boldsymbol{\mathcal{A}}$ coincide on the subspace $\bigcup_m \mathcal{L}_m$, where

$$\mathcal{L}_m = \{\phi \in L_{1,-1} : \phi \equiv 0 \text{ on } [m,\infty)\}. \tag{10}$$

Although our treatment of (5) was carried out in the space $L_{1,-1}$, we now choose the Banach space X defined by

$$X = \{\phi \in L_{1,-1} \cap L_1 \colon |||\phi||| = \|\phi\|_1 + \|\phi\|_{1,-1} < \infty\} \tag{11}$$

for our analysis of the full equation (2). This space provides a more convenient setting for establishing that the nonlinear coagulation operator N is locally Lipschitz. We obtain the local existence and uniqueness of a mild solution of (2) via a contraction mapping argument. Then, by applying [2, Thm. 2.6, pp. 90–91], we show that the solution is global.

In the first instance, the work presented here extends the results of Aizenman and Bak [1] for constant coagulation and fragmentation kernels to bounded coagulation and bounded, unsymmetric fragmentation kernels. However, in order to prove that the solutions are nonnegative, it has been necessary to restrict attention to constant coagulation kernels and bounded, unsymmetric fragmentation kernels.

Throughout, the following hypotheses are imposed on the kernels K and γ, with the further restriction, K constant, being added only as it is required.

The fragmentation kernel γ satisfies (H1), (H2), and

(H3)$'$ $\gamma \in L^\infty((0,\infty)\times(0,\infty))$.

The coagulation kernel K satisfies the following hypotheses:

(H4) $K(x,y)$ is a nonnegative function on $(0,\infty)\times(0,\infty)$;

(H5) K is symmetric, i.e., $K(x,y) = K(y,x)$ for all $x,y \in (0,\infty)$;

(H6) $K \in L^\infty((0,\infty)\times(0,\infty))$.

As a consequence of (H3)$'$ and (H6), for all $\Omega \subseteq (0,\infty)\times(0,\infty)$,

$$\operatorname*{ess\,sup}_{(x,y)\in\Omega} |\gamma(x,y)| \leq \frac{2}{3}C_F < \infty \qquad \text{where } \frac{2}{3}C_F = \|\gamma(x,y)\|_\infty \tag{12}$$

and

$$\operatorname*{ess\,sup}_{(x,y)\in\Omega} |K(x,y)| \leq C_K < \infty \qquad \text{where } C_K = \|K(x,y)\|_\infty.$$

(The factor 2/3 is chosen for its convenience in later calculations.) We note that hypothesis (H3) is automatically satisfied since

$$a(x) := \int_0^x \frac{y}{x}\gamma(x,y)\,dy \leq \frac{1}{3}C_F x \leq \frac{1}{3}C_F n \quad \text{for all } x \in (0,n].$$

Therefore, the analysis in [5] is also valid for fragmentation kernels γ satisfying (H1), (H2), and (H3)$'$. Also, it follows by [5, Thm. 5.6 and Ex. 5.7(i)] that whenever the fragmentation kernel γ satisfies (H3)$'$, the generator $\boldsymbol{\mathcal{A}}$ of the semigroup $\{S(t)\}_{t\geq 0}$ on $L_{1,-1}$ is a restriction of the operator A.

2. The pure fragmentation semigroup on X.

We begin by showing that the pure fragmentation semigroup $\{S(t)\}_{t\geq 0}$ introduced in (6) also forms a C$_0$-semigroup on X. We need only verify that

$$\lim_{t\to 0^+} |||S(t)\phi - \phi||| = 0 \qquad \text{for all } \phi \in X,$$

since $S(t+s) = S(t)S(s)$ for all $t,\ s \geq 0$, and $S(0) = I$ hold automatically.

Theorem 2.1. *For each $t \geq 0$, let $S(t)$ be the operator introduced in* (6). *Then $\{S(t)\}_{t\geq 0}$ forms a C_0-semigroup on X and*

$$|||S(t)\phi||| \leq (1 + tC_F)\,|||\phi||| \ \textit{ for all } \phi \in X. \tag{13}$$

Proof. Let $\phi \in \bigcup_m \mathcal{L}_m$, where $\mathcal{L}_m$ is defined by (10). By [5, Cor. 6.3], $S(t)\phi$ is a strong solution of the ACP (5) with initial data $f = \phi$, and hence

$$S(t)\phi = \phi + \int_0^t AS(\tau)\phi\,d\tau. \tag{14}$$

By Fubini's theorem and (H3)$'$, $\|[A\psi]\|_1 \leq C_F\,\|\psi\|_{1,-1}$ for all $\psi \in X$. Since the operators $\{S(t)\}_{t\geq 0}$ form a C$_0$-semigroup of isometries on $L_{1,-1}$ (by [5, Lem. 4.1]), it is then not difficult to show that for all $\phi \in \bigcup_m \mathcal{L}_m$,

$$|||S(t)\phi||| = \|S(t)\phi\|_1 + \|S(t)\phi\|_{1,-1} \leq |||\phi||| + \int_0^t \|[AS(\tau)\phi]\|_1\,d\tau \leq (1 + C_F t)\,|||\phi|||\,.$$

The latter result extends to all $\phi \in X$, since for each fixed $t \geq 0$, $S(t)$ is a bounded operator on X.

From (14), we also obtain that $\|S(t)\phi - \phi\|_1 \leq C_F t\,\|\phi\|_{1,-1}$ for $\phi \in \bigcup_m \mathcal{L}_m$. Therefore, for all $\phi \in \bigcup_m \mathcal{L}_m \subseteq X$, we deduce that $S(t)\phi \to \phi$ with respect to the $|||.|||$-norm as $t \to 0^+$. This result extends to all $\phi \in X$ by continuity since $\bigcup_m \mathcal{L}_m$ is dense in X and, for all $t \leq 1$, $\|S(t)\|_{B(X)} \leq (1 + C_F)$. □

It should be noted that the infinitesimal generator, $\mathbf{A}$, of $\{S(t)\}_{t\geq 0}$ on X is not necessarily the same as the infinitesimal generator, $\boldsymbol{\mathcal{A}}$, of $\{S(t)\}_{t\geq 0}$ on $L_{1,-1}$ because now

$$\operatorname*{s-lim}_{t\to 0^+} \frac{S(t)\phi - \phi}{t} \tag{15}$$

is with respect to the $|||.|||$-norm instead of with respect to the $\|.\|_{1,-1}$-norm. However, it is possible to prove that $\mathbf{A}$ is a restriction of $\boldsymbol{\mathcal{A}}$.

THEOREM 2.2. *Let* $\mathbf{A}$ *denote the generator of the semigroup* $\{S(t)\}_{t\geq 0}$ *on* X, *and let* $\boldsymbol{\mathcal{A}}$ *denote the generator of* $\{S(t)\}_{t\geq 0}$ *on* $L_{1,-1}$. *Then* $\mathbf{A}$ *is a restriction of* $\boldsymbol{\mathcal{A}}$ *and hence is also a restriction of the operator* A *defined by* (3).

Proof. Let $\phi \in \mathcal{D}(\mathbf{A})$. By definition, the limit (15) with respect to the $|||.|||$-norm exists and equals $\mathbf{A}\phi$. Hence the limit (15) with respect to the $\|.\|_{1,-1}$-norm exists and equals $\mathbf{A}\phi$. The latter implies that $\phi \in \mathcal{D}(\boldsymbol{\mathcal{A}})$ and $\boldsymbol{\mathcal{A}}\phi = \mathbf{A}\phi$. Thus $\mathbf{A}$ is a restriction of $\boldsymbol{\mathcal{A}}$, which in turn is a restriction of A by our earlier discussion. □

The previous result shows that we may write $\mathbf{A}\phi = A\phi$ for all $\phi \in \mathcal{D}(\mathbf{A})$. Moreover, the operator A restricted to $\mathcal{D}(\mathbf{A})$ generates the semigroup $\{S(t)\}_{t\geq 0}$ on X.

3. Existence and uniqueness of solutions to the coagulation and fragmentation equation.

Our first step is to establish that the operator N which appears in the inhomogeneous term $Nu(t)$ on the right-hand side of (2) is locally Lipschitz on X.

LEMMA 3.1. *Let the coagulation kernel* K *satisfy* (H4)–(H6). *Then the operator* $N : X \to X$ *defined by* (4) *is continuous and satisfies the Lipschitz condition*

$$|||N\phi - N\psi||| \leq 2C_K\,(|||\phi||| + |||\psi|||)\,|||\phi - \psi||| \tag{16}$$

whenever $\phi,\ \psi \in X$, *i.e.,* N *is locally Lipschitz on* X.

Proof. The proof is not difficult and makes use of Fubini's Theorem, (H6), and the simple result that $\|\phi\|_1 \leq |||\phi|||$ and $\|\phi\|_{1,-1} \leq |||\phi|||$ for all $\phi \in X$. See [4, Lem. 4.3, pp. 76–77] for details. □

To obtain the local existence and uniqueness of a mild solution to the initial-value problem (2), we use a contraction mapping argument in the Banach space $Y_{b,T}$ defined by

$$Y_{b,T} = \{y \in C([0,T], X) : |||y|||_\infty < b\}, \tag{17}$$

where

$$|||y|||_\infty = \sup_{t\in[0,T]} |||y(t)|||\,.$$

For given initial data $f \in X$, the constant b is chosen such that $|||f||| < b$, and then T is chosen sufficiently small in order to satisfy the inequalities that arise in subsequent proofs.

THEOREM 3.2. *Let* $f \in X$. *For a suitable choice of* T *and* b, *the map* $u \to W_f u$ *defined by*

$$[W_f u](t) = S(t)f + \int_0^t S(t-\sigma)N(u(\sigma))\,d\sigma, \quad t \in [0,T], \tag{18}$$

has a unique fixed point $\tilde{u}$ *in the space* $Y_{b,T}$ *defined by* (17). *Moreover,* $\tilde{u}(0) = f$.

Proof.

(i) Our first step is to prove that $W_f(Y_{b,T}) \subseteq Y_{b,T}$. Let $u \in Y_{b,T}$. Then, for all $t \in [0,T]$, $N(u(t)) \in X$ and

$$|||N(u(t)))||| \leq 2C_K\,|||u(t)|||^2 \leq 2C_K b^2 < \infty,$$

by (16) with $\phi = u(t)$ and $\psi \equiv 0$. Therefore, using (13), we obtain

$$|||[W_f u](t)||| \leq |||f||| + TC_F b + \left(T + \frac{T^2}{2}C_F\right)2C_K b^2. \tag{19}$$

Since the right-hand side of (19) is finite, $[W_f u](t) \in X$ for all $t \in [0,T]$. Now let $t_1, t_2 \in [0,T]$ with $t_1 = t_2 + t,\ t > 0$. Then

$$\begin{aligned}&|||[W_f u](t_1) - [W_f u](t_2)|||\\ &\le |||(S(t)-I)S(t_2)f||| + \int_{t_2}^{t_1} |||S(t_1 - t_2)S(t_2 - \sigma)N(u(\sigma))|||\, d\sigma\\ &\qquad\qquad + \int_0^{t_2} |||(S(t) - I)S(t_2 - \sigma)N(u(\sigma))|||\, d\sigma.\end{aligned}$$

Since $S(t)$ is strongly continuous on X (by Theorem 2.1), we deduce that $W_f u \in C([0,T], X)$. Moreover, using (19) and choosing T sufficiently small, we arrive at

$$|||W_f u|||_\infty \le |||f||| + TC_F b + \left(T + \frac{T^2}{2} C_F\right) 2C_K b^2 < b.$$

(Note that such a choice of T is possible by continuity since the strict inequality holds for $T = 0$ and hence for T sufficiently close to zero.) Hence $|||W_f u|||_\infty < b$, and so $W_f u \in Y_{b,T}$ whenever $u \in Y_{b,T}$.

(ii) We now show that the map $u \to W_f u$ is a contraction mapping on $Y_{b,T}$. Let $u, v \in Y_{b,T}$. By (13), (16), and (18),

$$\begin{aligned}|||W_f u - W_f v|||_\infty &\le \sup_{t \in [0,T]} 4C_K b \int_0^t (1 + (t-\sigma)C_F)\, |||u(\sigma) - v(\sigma)|||\, d\sigma\\ &\le 4C_K b \left(T + \frac{T^2}{2} C_F\right) |||u - v|||_\infty .\end{aligned}$$

We now choose T sufficiently small so that, in addition to the previous requirement on T,

$$|||W_f u - W_f v|||_\infty \le \lambda\, |||u - v|||_\infty \quad \text{for some constant } \lambda \in (0,1);$$

i.e., we choose T such that

$$4C_K b \left(T + \frac{T^2}{2} C_F\right) \le \lambda$$

is also valid. Then, by the contraction mapping principle, there exists a unique fixed point $\tilde{u} \in Y_{b,T}$. Finally, we note that $\tilde{u}(0) = [W_f \tilde{u}](0) = f$. □

THEOREM 3.3. *Let u be the unique fixed point of the map $u \to W_f u$ obtained in Theorem 3.2. Then*

$$\int_0^\infty x[u(t)](x)\, dx = \int_0^\infty x f(x)\, dx \qquad \textit{for all } t \in [0,T].$$

Proof. By definition,

$$u(t) = S(t)f + \int_0^t S(t-\sigma)N(u(\sigma))\, d\sigma, \qquad 0 \le t \le T.$$

Therefore, by (7) and (8) (see [5, Lem. 4.1(ii)]),

$$\int_0^\infty x[u(t)](x)\, dx = \int_0^\infty x f(x)\, dx + I(t),$$

where

$$I(t) = \int_0^\infty x \left[\int_0^t S(t-\sigma) N(u(\sigma))\, d\sigma \right](x)\, dx.$$

Thus, it remains to prove that $I(t) = 0$ for all $t \in [0, T]$.

It follows from Theorem 2.1 and Lemma 3.1 that $N(u(.))$ and $S(t-.)N(u(.))$ are strongly continuous on X. Since X is a Banach space of type L [3, pp. 69–71], we note that

$$\left[\int_0^t S(t-\sigma) N(u(\sigma))\, d\sigma \right](x) = \int_0^t \left[S(t-\sigma) N(u(\sigma)) \right](x)\, d\sigma.$$

By Fubini's theorem, (7), and (8),

$$I(t) = \int_0^t \int_0^\infty x[N(u(\sigma))](x)\, dx\, d\sigma.$$

By interchanging orders of integration and making a change of variable $x' = x - y$ on the left-hand side, it is not difficult to show that

$$\frac{1}{2} \int_0^\infty \int_0^x xK(x-y,y)\phi(x-y)\phi(y)\, dy\, dx = \int_0^\infty x\phi(x) \int_0^\infty K(x,y)\phi(y)\, dy\, dx,$$

and so from (4),

$$\int_0^\infty x[N(u(\sigma))](x)\, dx = 0, \qquad 0 \le \sigma \le t \le T. \qquad \square$$

THEOREM 3.4 (existence and uniqueness of a mild solution). *Let $f \in X$. Then there exists $T \in (0, \infty]$ such that for $t \in [0, T)$ equation* (2) *has a unique mild solution with $u(0) = f$ and $u(t) \in X$ for all $t \in [0, T)$. The maximal T, $\widehat{T}$, with this property is finite only if* $\lim_{t \to \widehat{T}^-} |||u(t)||| = \infty$.

Proof. The mild solution of equation (2) on $[0, T]$ is given by the unique fixed point of the map $u \to W_f u$ obtained in Theorem 3.2. The result follows directly from [2, Thm. 2.6, pp. 90–91]. $\square$

COROLLARY 3.5. *Let u be the mild solution of* (2) *on* $[0, \widehat{T})$. *Then*

$$\int_0^\infty x[u(t)](x)\, dx = \int_0^\infty xf(x)\, dx \qquad \textit{for all } t \in [0, \widehat{T}).$$

Proof. The result follows from Theorems 3.3 and 3.4. $\square$

4. A nonnegative solution to the coagulation and fragmentation equation. In this section we prove that the mild solution obtained in Theorem 3.4 is nonnegative whenever the initial data f is nonnegative. This result is analogous to that given in [1] for the case of constant fragmentation kernels and is established by a similar argument. However, as the justification provided in [1] is somewhat incomplete, we feel it is desirable to include more detail here.

We now assume that the coagulation kernel K satisfies

(H7) $K(x, y) = C_K > 0$ for all $(x, y) \in (0, \infty) \times (0, \infty)$, in which case (H4)–(H6) are immediately satisfied. The operator N defined by (4) now reduces to

$$[N\phi](x) = \frac{C_K}{2}\int_0^x \phi(x-y)\phi(y)\,dy - C_K\phi(x)\int_0^\infty \phi(y)\,dy, \tag{20}$$

and the map $u(t) \to [W_f u](t)$, defined by (18), becomes

$$u(t) \to S(t)f + \int_0^t S(t-\sigma)\left\{\frac{C_K}{2}\int_0^{\cdot}[u(\sigma)](.-y)[u(\sigma)](y)\,dy \right. \\ \left. -C_K[u(\sigma)]\int_0^\infty [u(\sigma)](y)\,dy\right\}d\sigma. \tag{21}$$

In the previous section we proved, via a contraction mapping argument on the space $Y_{b,T}$ (defined by (17)), that the map $u \to W_f u$ has a unique fixed point $\tilde{u} \in Y_{b,T}$. The contraction mapping theorem also tells us that for any initial estimate $u_{(1)} \in Y_{b,T}$, the successive approximations $u_{(n+1)} = W_f u_{(n)}$ $(n = 1, 2, \dots)$ converge to $\tilde{u}$.

We assume that $\tilde{u}$ is known and consider the successive approximation scheme

$$u_{(n+1)}(t) = S(t)f + \int_0^t S(t-\sigma)\left\{\frac{C_K}{2}\int_0^{\cdot}[\tilde{u}(\sigma)](.-y)[\tilde{u}(\sigma)](y)\,dy \right. \\ \left. + C_K\left(-\int_0^\infty [\tilde{u}(\sigma)](y)\,dy\right)u_{(n)}(\sigma)\right\}d\sigma, \tag{22}$$

$n = 1, 2, \dots$, with a suitable $u_{(1)}(t)$. Notice that we substitute into only one term. We note that $\tilde{u}$ is a fixed point of the associated map

$$\phi(t) \to S(t)f + \int_0^t S(t-\sigma)\left\{\frac{C_K}{2}\int_0^{\cdot}[\tilde{u}(\sigma)](.-y)[\tilde{u}(\sigma)](y)\,dy \right. \\ \left. + C_K\left(-\int_0^\infty [\tilde{u}(\sigma)](y)\,dy\right)\phi(\sigma)\right\}d\sigma. \tag{23}$$

We can show that the map (23) is also a contraction mapping on the space $Y_{b,T}$, and so $\tilde{u}$ is the unique fixed point in $Y_{b,T}$. From the contraction mapping principle applied to the map (23) we obtain that the successive approximations (22) converge to $\tilde{u}$.

We now use the successive approximation scheme (22) to prove that $\tilde{u}$ is nonnegative. We begin by introducing some additional notation to ease computation.

NOTATION 4.1. *Let $\{S(t)\}_{t\geq 0}$ be the C_0-semigroup on X introduced in* (6). *Let $\tilde{u}$ be the unique fixed point of the map $u \to W_f u$ (defined by* (21)*) and suppose that $\tilde{u}$ is known. Then we define*

(i) *$h\colon [0,T] \to X$ by $h(t) := S(t)f$, $t \in [0,T]$;*

(ii) *$g\colon (t,\sigma) \to g(t,\sigma) \in X$, $0 \leq \sigma \leq t \leq T$, by*

$$g(t,\sigma) := S(t-\sigma)\left[\frac{C_K}{2}\int_0^{\cdot}[\tilde{u}(\sigma)](.-y)[\tilde{u}(\sigma)](y)\,dy\right] = S(t-\sigma)\mathcal{N}_1(\tilde{u}(\sigma)),$$

where $\mathcal{N}_1$ is defined for $\phi \in X$ by

$$(\mathcal{N}_1\phi)(x) := \frac{C_K}{2}\int_0^x \phi(x-y)\phi(y)\,dy, \qquad \text{a.e. } x > 0; \tag{24}$$

(iii) *$\xi\colon [0,T] \to \mathbf{R}$ by $\xi(t) := -C_K\int_0^\infty [\tilde{u}(t)](y)\,dy$;*

(iv) $\beta\colon [0,T]\times[0,T]\to \mathbf{R}$ *by* $\beta(t,\sigma) := \int_\sigma^t \xi(\tau)\,d\tau$*;*

(v) $\alpha : [0,T] \to \mathbf{R}$ *by* $\alpha(t) := \beta(t,0)$*.*

On using the above notation, we can write the successive approximation scheme (22) as

$$u_{(n+1)}(t) = h(t) + \int_0^t g(t,\sigma)\,d\sigma + \int_0^t \xi(\sigma)S(t-\sigma)u_{(n)}(\sigma)\,d\sigma, \tag{25}$$

$n = 1, 2, \ldots$. We can take $\xi(s)$ (also $\alpha(t)$ and $\beta(t,\sigma)$) outside the operator $S(t)$, $t \geq 0$, because the operators $\{S(t)\}_{t\geq 0}$ are linear and $\xi(\sigma)$, $\alpha(t)$, and $\beta(t,\sigma)$ are real numbers.

Lemma 4.2. *For all* $0 \leq r \leq \sigma \leq t \leq T$*,*

$$S(t-\sigma)h(\sigma) = h(t) \text{ and } S(t-\sigma)g(\sigma,r) = g(t,r).$$

Proof. This follows from the semigroup properties of $\{S(t)\}_{t\geq 0}$; see [4, Lem. 4.9, p. 86] for details. □

Lemma 4.3. *For* $0 \leq \sigma \leq t \leq T$, $n \in \mathbf{N}$*,*

$$\int_\sigma^t \xi(\tau)\frac{\beta(\tau,\sigma)^n}{n!}\,d\tau = \frac{\beta(t,\sigma)^{n+1}}{(n+1)!}.$$

Proof. This is a simple proof by induction. See [4, Lem. 4.10, pp. 86–87] for details. □

Lemma 4.4. *For all* $t \in [0,T]$ *and* $n \in \mathbf{N}$*,*

$$\int_0^t \xi(\sigma)\int_0^\sigma g(t,\tau)\frac{\beta(\sigma,\tau)^n}{n!}\,d\tau\,d\sigma = \int_0^t g(t,\sigma)\frac{\beta(t,\sigma)^{n+1}}{(n+1)!}\,d\sigma.$$

Proof. This follows from Lemma 4.3 on changing the order of integration. □

Lemma 4.5. *Let* $u_{(1)}(t)$ *be defined by*

$$u_{(1)}(t) = h(t) + \int_0^t g(t,\sigma)\,d\sigma, \qquad t \in [0,T]. \tag{26}$$

Then $u_{(1)} \in Y_{b,T}$*.*

Proof. The details are similar to the proof of Theorem 3.2, part (i). See [4, Lem. 4.13, pp. 87–89] for details. □

Lemma 4.6. *Let* $u_{(1)}(t)$ *be defined by* (26)*. Then*

$$u_{(n)}(t) = \left[\sum_{i=0}^{n-1}\frac{\alpha(t)^i}{i!}\right]h(t) + \int_0^t \left[\sum_{i=0}^{n-1}\frac{\beta(t,\sigma)^i}{i!}\right]g(t,\sigma)\,d\sigma, \quad n = 1, 2, 3, \ldots,$$

is the general formula for the nth iterate of the successive approximation scheme (25)*.*

Proof. This proof by induction uses Lemma 4.3, with $\sigma = 0$, and Lemma 4.4. We note that we can take $S(t-\sigma)$ under the integral sign because it is a bounded linear operator on X. □

Lemma 4.7. *The fixed point* $\tilde{u}$ *of the map* $u \to W_f u$*, where* $W_f u$ *is given by* (21)*, also satisfies*

$$\tilde{u}(t) = [\exp \alpha(t)]h(t) + \int_0^t [\exp \beta(t,\sigma)]g(t,\sigma)\,d\sigma, \quad t \in [0,T]. \tag{27}$$

(Recall that $\alpha(t), \beta(t,\sigma)$, and $g(t,\sigma)$ depend on $\tilde{u}$.)

Proof. This follows immediately by letting $n \to \infty$ in Lemma 4.6. ◻

The mild solution $u(t)$ of (2) (with $K(x,y) = C_K$) can be obtained as the fixed point of the map $u \to W_f u$ or as the fixed point of the map (23). The above lemma suggests that this mild solution can also be obtained as a fixed point of the map $u \to Vu$ on $Y_{b,T}$ defined by

$$(28) \qquad (Vu)(t) = [\exp \alpha(t)]h(t) + \int_0^t [\exp \beta(t,\sigma)]S(t-\sigma)[\mathcal{N}_1(u(\sigma))]\, d\sigma$$

where $\mathcal{N}_1(u(\sigma))$ is given by (24). The map $u \to Vu$ is obtained from (27) by keeping $\tilde{u}$ in the definition of $\alpha(t)$ and $\beta(t,\sigma)$ but not so in the definition of $g(t,\sigma)$. Therefore, now $g(t,\sigma) = S(t-\sigma)\mathcal{N}_1(u(\sigma))$.

LEMMA 4.8. *For a suitable choice of T and b, the map $u \to Vu$ defined by* (28) *is a contraction mapping on the space $Y_{b,T}$ (defined by* (17)*) and therefore has a unique fixed point $\hat{u} \in Y_{b,T}$. Moreover, for suitable T and b, the fixed point $\hat{u}$ of the map $u \to Vu$ coincides with the fixed point $\tilde{u}$ of the map $u \to W_f u$.*

Proof. As the proof is analogous to the proof of Theorem 3.2, we provide only a brief outline. For further details see [4, Lem. 4.16, pp. 90–93].

(i) We note that for all $u \in Y_{b,T}$, $|\xi(\tau)| \le C_K\, |||\tilde{u}(\tau)||| < C_K b$. Hence, for all $t, \sigma \in [0,T]$, $\exp \beta(t,\sigma) \le \exp(C_K bT)$, and $\exp \alpha(t) \le \exp(C_K bT)$. So for all $t \in [0,T]$, we find

$$(29) \qquad |||(Vu)(t)||| \le \exp(C_K bT)\left[(1+TC_F)\, |||f||| + \left(T + \frac{T^2}{2}C_F\right) C_K b^2\right].$$

To prove $(Vu)(t)$ is a continuous function of t we let $t_1,\ t_2 \in [0,T]$, with $t_1 > t_2$, and consider

$$(30) \qquad |||(Vu)(t_1) - (Vu)(t_2)|||\,.$$

Now (30) can be shown to be bounded by

$$\begin{aligned} &|||\{\exp \alpha(t_1)\}h(t_1) - \{\exp \alpha(t_2)\}h(t_2)||| \\ &\qquad + \int_{t_2}^{t_1} \exp \beta(t_1,\sigma)\, |||S(t_1-t_2)S(t_2-\sigma)\mathcal{N}_1(u(\sigma))|||\, d\sigma \\ &\qquad\qquad - \int_0^{t_2} |||\{\exp \beta(t_1,\sigma)S(t_1-t_2) - \exp \beta(t_2,\sigma)\}S(t_2-\sigma)\mathcal{N}_1(u(\sigma))|||\, d\sigma, \end{aligned}$$

and since

$$|||\exp \beta(t_1,t_2)S(t_1-t_2)f - f||| \to 0 \text{ as } t_1 \to t_2^+,$$

for any $f \in X$, the required continuity of $(Vu)(t)$ follows.

(ii) Using (29) and choosing T appropriately, we obtain

$$|||(Vu)(t)||| \le \exp(C_K bT)(1+TC_F)\, |||f||| + \exp(C_K bT)\left(T + \frac{T^2}{2}C_F\right) C_K b^2 < b$$

for all $t \in [0,T]$, and so $V(Y_{b,T}) \subseteq Y_{b,T}$.

(iii) We can show that for $u, v \in Y_{b,T}$, $\lambda \in (0,1)$,

$$|||(Vu)(t) - (Vv)(t)||| \leq 2C_K b \exp(C_K bT) \left(T + \frac{T^2}{2} C_F\right) |||u - v|||_\infty \leq \lambda \, |||u - v|||_\infty ,$$

provided T is such that

$$2C_K b \exp(C_K bT) \left(T + \frac{T^2}{2} C_F\right) \leq \lambda.$$

(iv) We recall that the map $u \to Vu$ is obtained from (27) by replacing $\tilde{u}$ by u in the definition of $g(t, \sigma)$. Therefore $\tilde{u}$ is a fixed point of $u \to Vu$. Since the fixed point of the map $u \to Vu$ in the space $Y_{b,T}$ is unique, it follows that $\tilde{u} \equiv \hat{u}$. □

The above lemma enables us to state that, for a suitable choice of T and any $u_{(1)} \in Y_{b,T}$, the successive approximations $u_{(n+1)} = Vu_{(n)}$ converge to the fixed point of the map $u \to W_f u$.

Theorem 4.9 (nonnegative solution). *The mild solution u of* (2) *on* $[0, T]$ *is nonnegative whenever the initial data $f \in X$ is nonnegative.*

Proof. From the above discussion and Theorem 3.4, the mild solution u can be found from the nonnegative preserving iterations $u_{(n+1)} = Vu_{(n)}$, $u_{(1)} = f$. □

5. A global solution. By Theorem 3.4, we have the existence and uniqueness of a global mild solution, u, unless there exists a $\widehat{T} < \infty$, such that

$$\lim_{t \to \widehat{T}^-} |||u(t)||| = \infty.$$

We shall prove that there is no such finite $\widehat{T}$ and so establish that there exists a unique, global, nonnegative, mild solution to the ACP formulation (2) of the full coagulation and fragmentation equation (1) with K constant and γ bounded.

Let u be the mild solution of (2) on $[0, \widehat{T})$ and consider the functions M_i $(i = 0, 1)$ defined by

$$M_i(t) = \int_0^\infty x^i [u(t)](x) \, dx, \quad t \in [0, \widehat{T}). \tag{31}$$

For nonnegative $u(t)$,

$$|||u(t)||| = M_0(t) + M_1(t).$$

By Corollary 3.5,

$$M_1(t) = \int_0^\infty x f(x) \, dx \qquad \text{for all } t \in [0, \widehat{T}),$$

and so it follows that

$$\lim_{t \to \widehat{T}^-} |||u(t)||| = \infty \qquad \text{only if} \qquad \lim_{t \to \widehat{T}^-} M_0(t) = \infty. \tag{32}$$

To prove that $\lim_{t \to \widehat{T}^-} M_0(t) = \infty$ cannot hold for any finite $\widehat{T}$, we make use of the following proposition.

PROPOSITION 5.1. *Let $u(t)$ be the mild solution of* (2) *on* $[0, \widehat{T})$. *Then* $u(x,t) = [u(t)](x)$ *satisfies*

$$\begin{aligned} u(x,t) =& f(x) + \int_0^t \left\{ \int_x^\infty \gamma(y,x)u(y,\sigma)\,dy - u(x,\sigma)\int_0^x \frac{y}{x}\gamma(x,y)\,dy \right. \\ &\left. + \frac{1}{2}\int_0^x K(x-y,y)u(x-y,\sigma)u(y,\sigma)\,dy - u(x,\sigma)\int_0^\infty K(x,y)u(y,\sigma)dy \right\} d\sigma \end{aligned}$$

for all $t \in [0, \widehat{T})$, *a.e.* $x > 0$.

Proof. The argument is somewhat involved because we must consider a truncated version of the full equation, even though this step was not necessary to prove existence and uniqueness of a local solution. We postpone this proof until the end of section 6. □

THEOREM 5.2. *Let* $M_0(t)$ *be defined by* (31) *(with* $i = 0$*) and suppose that for all* $t \in [0, \widehat{T})$, $u(t) \geq 0$. *Then*

$$M_0(t) \leq \int_0^\infty f(x)\,dx + \frac{2}{3}tC_F \int_0^\infty xf(x)\,dx.$$

Proof. For all $t \in [0, \widehat{T})$, by Proposition 5.1 and repeated use of Fubini's theorem,

$$\begin{aligned} \int_0^\infty [u(t)](x)\,dx &= \int_0^\infty f(x)\,dx + \int_0^t \left[-\frac{1}{2}\int_0^\infty \int_0^\infty K(x,y)u(y,\sigma)u(x,\sigma)\,dy\,dx \right. \\ &\qquad \left. + \int_0^\infty \int_0^x \left(1 - \frac{y}{x}\right)\gamma(x,y)\,dy\,u(x,\sigma)\,dx \right] d\sigma \\ &\leq \int_0^\infty f(x)\,dx + \int_0^t \int_0^\infty \int_0^x \gamma(x,y)\,dy\,u(x,\sigma)\,dx\,d\sigma \\ &\qquad \text{(since } u(\sigma) \geq 0 \text{ for all } \sigma \in [0,t],\ t \in [0,\widehat{T})\ \text{)} \\ &\leq \int_0^\infty f(x)\,dx + \int_0^t \int_0^\infty \frac{2}{3}C_F x u(x,\sigma)\,dx\,d\sigma \quad \text{(by (12))} \\ &= \int_0^\infty f(x)\,dx + \frac{2}{3}tC_F \int_0^\infty xf(x)\,dx. \qquad \square \end{aligned}$$

COROLLARY 5.3. *There does not exist a finite* $\widehat{T}$ *such that*

$$\lim_{t \to \widehat{T}^-} |||u(t)||| = \infty.$$

Proof. This follows immediately from Theorem 5.2 and (32). □

Thus, by the above discussion, we have established the following theorem.

THEOREM 5.4. *Let the fragmentation kernel* γ *satisfy* (H1), (H2), *and* (H3)′, *and let the coagulation kernel* K *satisfy* (H7). *Then the ACP* (2) *has a unique, global, nonnegative, mass-conserving mild solution.*

We remark that Proposition 5.1 enables us to relate the mild solution of the ACP (2) to a solution of the coagulation and fragmentation equation (1).

THEOREM 5.5. *Let* γ *and* K *be as in Theorem* 5.4. *Then the coagulation and fragmentation equation* (1) *has a nonnegative mass-conserving solution* u *given by*

$$u(x,t) = [u(t)](x) \qquad \text{for a.e. } (x,t) \in (0,\infty) \times [0,\infty),$$

where $u(t)$ $(t \geq 0)$ is the unique mild solution of the ACP (2) *obtained in Theorem* 5.4. *Moreover, this is the only solution of* (1) *such that the vector-valued function $u(.,t)$, $t \geq 0$, is a mild solution of the ACP* (2).

Proof. By Proposition 5.1, the scalar-valued function u, constructed via

$$u(x,t) = [u(t)](x) \qquad \text{for all } t \geq 0 \text{ and a.e. } x \in (0,\infty),$$

is absolutely continuous and satisfies (1) for a.e. $(x,t) \in (0,\infty) \times [0,\infty)$. The scalar-valued function inherits the nonnegativity and mass conservation from the vector-valued function. Uniqueness of the mild solution of (2) implies the uniqueness statement in the theorem. $\square$

6. A truncated coagulation and fragmentation equation. We now investigate a truncated version of the full coagulation and fragmentation equation (1) and an associated truncated version of the ACP (2). The truncated coagulation and fragmentation equation is given by

(33)
$$\frac{\partial}{\partial t}u(x,t) = \begin{cases} \frac{1}{2}\int_0^x K(x-y,y)u(x-y,t)u(y,t)dy - u(x,t)\int_0^{m-x} K(x,y)u(y,t)dy \\ \qquad + \int_x^m \gamma(y,x)u(y,t)dy - a(x)u(x,t), & \text{for a.e. } x < m, \\ \\ 0, & \text{for a.e. } x \geq m, \end{cases}$$

where a is defined by (9). We make use of the projection operator P_m $(m > 0)$ on X defined by

$$(P_m\phi)(x) = \begin{cases} \phi(x), & 0 < x < m, \\ 0, & x \geq m. \end{cases}$$

We seek a sequence $\{u_m(t)\}_{m>0}$ of approximating solutions satisfying the truncated ACP,

$$\begin{aligned} \frac{d}{dt}u_m(t) &= A_m u_m(t) + N_m u_m(t), \quad t > 0, \\ u_m(0) &= f_m, \end{aligned} \tag{34}$$

where $A_m \equiv AP_m$ and where

(35)
$$(N_m\phi)(x) = \begin{cases} \frac{1}{2}\int_0^x K(x-y,y)\phi(x-y)\phi(y)dy - \phi(x)\int_0^{m-x} K(x,y)\phi(y)dy, & x<m, \\ 0, & x \geq m. \end{cases}$$

Note that $N_m\phi \not\equiv NP_m\phi$. The initial data f_m for the truncated problem is chosen so that $|||f_m - f||| \to 0$ as $m \to \infty$, where f is the initial data for the nontruncated problem in X. At this point, we do not state the form that f_m takes, but note that f_m could equal $P_m f$.

For each fixed $m > 0$, a mild solution u_m of (34) satisfies the integral equation

$$u_m(t) = S_m(t)f_m + \int_0^t S_m(t-s)N_m(u_m(s))\,ds, \quad t \geq 0, \tag{36}$$

where $\{S_m(t)\}_{t\geq 0}$ is the linear semigroup generated by A_m on the space X defined by (11). So, we consider the map $u \to W_{m,f_m}u$ where, for each fixed $t \in [0,T]$,

$$[W_{m,f_m}u](t) = S_m(t)f_m + \int_0^t S_m(t-s)[(N_m u)(s)]\,ds. \tag{37}$$

(Compare (37) with the definition of $W_f u$ given by (18).) By repeating the argument used for the map $u \to W_f u$ in sections 2 and 3, we can prove the existence and uniqueness of a mild solution of (34). The appropriate norm estimates for the truncated operators are stated in the next two lemmas.

Lemma 6.1. *Let $\{S(t)\}_{t\geq 0}$ and $\{S_m(t)\}_{t\geq 0}$ be the linear semigroups generated by A and A_m, respectively, on X. Then for all $m > 0$, $t \geq 0$, and $\phi \in X$,*

(i) $S_m(t)\phi = (I - P_m)\phi + S(t)P_m\phi$;

(ii) $|||S_m(t)\phi||| \leq (1 + tC_F)\,|||\phi|||$;

(iii) $|||S(t)\phi - S_m(t)\phi||| \leq (2 + tC_F)\,|||\phi|||$.

Proof. Part (i) is a rewrite of [5, Lem. 4.1(iii)]. For parts (ii) and (iii), we use (12) in conjunction with part (i). □

Lemma 6.2. *The operator $N_m\colon X \to X$, defined by* (35), *is continuous and satisfies the Lipschitz condition*

$$|||N_m\phi - N_m\psi||| \leq 2C_K\,(|||\phi||| + |||\psi|||)\,|||\phi - \psi||| \tag{38}$$

whenever ϕ, $\psi \in X$. (Compare (38) *with the inequality* (16) *for N.)*

Proof. The proof is almost identical to that for Lemma 3.1. See [4, Lem. 4.24, pp. 99–100] for details. □

The next theorem is analogous to the combined results of Theorems 3.2 and 3.4.

Theorem 6.3. *For each fixed $m > 0$, let $f_m \in X$. Then, for a suitable choice of T and b, the map $u \to W_{m,f_m}u$ has a unique fixed point u_m in the space $Y_{b,T}$ defined by* (17), *where now the choice of b is such that $b > \max(|||f_m|||\,, |||f|||)$. Furthermore, there exists $T \in (0, \infty]$ such that the ACP* (34) *has a unique mild solution $u_m(t) \in X$ for all $t \in [0, T)$. The maximal T, $\widehat{T}$, with this property is finite only if $\lim_{t\to\widehat{T}^-} |||u_m(t)||| = \infty$.*

Proof. The argument follows from that outlined in the proof of Theorem 3.2 for the fixed point of the map $u \to W_f u$. We replace u, $S(t)$, Nu, and $W_f u$ by u_m, $S_m(t)$, $N_m u_m$, and $W_{m,f_m}u$, respectively. Note that $S_m(t)$ is uniformly continuous with respect to the $|||.|||$-norm since the generator A_m of $\{S_m(t)\}_{t\geq 0}$ is a bounded linear operator on X. □

Remark 6.4. The truncated data f_m is chosen such that $f_m \to f$ in X. Therefore, there exists $b > 0$ such that $b > \max(|||f_m|||\,, |||f|||)$ for all $m > 0$, and it may be assumed that the choice of b and T in the above theorem coincides with the choice of b and T in Theorem 3.2. Thus the contraction constant $\lambda \in (0, 1)$ such that

$$|||W_{m,f_m}u - W_{m,f_m}v||| \leq \lambda\,|||u - v|||$$

is independent of m.

Corollary 6.5. *The mild solution u_m of* (34) *obtained in Theorem* 6.3 *is also a strong solution.*

Proof. Note that A_m is a bounded operator on X. Also, $S_m(t - .)N_m(u_m(.))$ can be shown to be strongly continuous. Therefore, differentiating the integral equation (36), we see that the mild solution also satisfies (34) in a strong sense. □

Corollary 6.6. *Let $u_m(x, t) = [u_m(t)](x)$ for a.e. $x > 0$, $t \in [0, T]$. Then for all $t > 0$, $u_m(\,.\,,\,.\,)$ satisfies the truncated coagulation and fragmentation equation* (33) *a.e. and $u_m(x, 0) = f_m(x)$ a.e.*

Proof. By Corollary 6.5, $u_m(t)$ is a strong solution of (34). The proof now follows along similar lines to the proof of [5, Thm. 7.1]. Since X is a Banach space of type L

[3, pp. 69–70], for $t \in (0,T]$ and a.e. $x \in (0,\infty)$,

$$\frac{\partial}{\partial t}u_m(x,t) = \left[\frac{d}{dt}u_m(t)\right](x) = [A_m u_m(t)](x) + [N_m u_m(t)](x),$$

and the right-hand side of the latter equals the right-hand side of (33), as required. ☐

We can use this pointwise version of the solution to the truncated problem to prove Proposition 5.1. We shall require the following two lemmas.

Lemma 6.7. *Let $\phi \in X$ and let the operators N and N_m be defined by* (4) *and* (35)*, respectively. Then*

$$|||N_m\phi - N\phi||| \to 0, \qquad \text{as } m \to \infty.$$

Proof. Let $v_m(x) = (N_m\phi)(x) - (N\phi)(x)$, $x > 0$. For a given $x > 0$, we can always choose m sufficiently large so that

$$v_m(x) = \phi(x)\int_{m-x}^{\infty} K(x,y)\phi(y)\,dy.$$

Since $K(x,y) \le C_K$ and $\phi \in X \subseteq L_1$, it follows that $v_m(x) \to 0$ for a.e. $x > 0$ as $m \to \infty$. Also, for all $x > 0$,

$$|(1+x)v_m(x)| \le C_K\,|||\phi|||\,(1+x)|\phi(x)| + \frac{C_K}{2}(1+x)\int_0^x |\phi(x-y)||\phi(y)|\,dy,$$

and the right-hand side is an integrable function over $(0,\infty)$. By the Lebesgue dominated convergence theorem,

$$\int_0^\infty (1+x)|v_m(x)|\,dx \to \int_0^\infty (1+x).0\,dx = 0 \quad \text{as } m \to \infty;$$

i.e., $|||N_m u - Nu||| \to 0$ as $m \to \infty$. ☐

Lemma 6.8. *Let $f_m \in \mathcal{L}_m$. For a suitable choice of T and b, let u be the mild solution of* (2) *on $[0,T]$, and let u_m be the mild solution of* (34) *on the same time-interval. Then, for all $t \in [0,T]$,*

$$|||u_m(t) - u(t)||| \to 0 \qquad \textit{whenever} \qquad |||f_m - f||| \to 0,$$

and the convergence is uniform in t.

Proof. We choose T so that the contraction mapping principle can be applied simultaneously to the maps $u \to W_f u$ and $u \to W_{m,f_m}u$ on the space $Y_{b,T}$ defined by (17). (See Theorems 3.2 and 6.3.) Such a choice of T is possible by Remark 6.4. The contraction mapping principle tells us that

$$|||W^k_{m,f_m}u - u_m|||_\infty \to 0 \qquad \text{as } k \to \infty,$$

since the mild solution $u \in Y_{b,T}$ is a suitable choice for the initial estimate. Using a telescoping series we obtain

$$|||u_m - u|||_\infty \le \sum_{k=0}^{\infty} \lambda^k\,|||W_{m,f_m}u - u|||_\infty$$

where the contraction constant $\lambda < 1$ is independent of m (Remark 6.4). Hence

$$|||u_m - u|||_\infty \leq (1-\lambda)^{-1} \, |||W_{m,f_m} u - u|||_\infty .$$

Now

$$\begin{aligned} &|||[W_{m,f_m} u](t) - u(t)||| \\ &\leq |||S_m(t) f_m - S(t) f||| + \int_0^t |||S_m(t-s)[N_m(u(s))] - S(t-s)[N(u(s))]||| \, ds \\ &= |||S(t)[f_m - f]||| + \int_0^t |||S(t-s)[N_m(u(s)) - N(u(s))]||| \, ds, \end{aligned}$$

since for all $t \in [0,T]$, $N_m(u(t)) \in \mathcal{L}_m$ (see (35)) and $S(t)\phi = S_m(t)\phi$ whenever $\phi \in \mathcal{L}_m$ (from Lemma 6.1(i)). Therefore, by (12),

$$|||W_{m,f_m} u - u|||_\infty \leq (1+TC_F) \, |||f_m - f||| + (1+TC_F) \int_0^T |||N_m(u(s)) - N(u(s))||| \, ds.$$

By the bounded convergence theorem and Lemma 6.7,

$$\int_0^T |||N_m(u(s)) - N(u(s))||| \, ds \to 0 \quad \text{as} \quad m \to \infty.$$

Hence, finally, as $m \to \infty$,

$$|||u_m - u|||_\infty \to 0 \quad \text{whenever} \quad |||f_m - f||| \to 0. \qquad \square$$

Proof of Proposition 5.1. The idea behind this proof is similar to that of [5, Thm. 7.3]. We know that the solution $u_m(t)$ of the truncated problem converges strongly in X, uniformly with respect to $t \in [0,T]$, to the solution $u(t)$ of the full equation whenever the truncated data $f_m \in \bigcup_n \mathcal{L}_n$ converges strongly to f. Therefore, repeating the argument used in the proof of [5, Thm. 7.3], we obtain the existence of a subsequence $\{m_k\}$ such that whenever $f_{m_k} \to f$ strongly in X as $m_k \to \infty$, $u_{m_k}(x,t) \to u(x,t)$ for all $t \in [0,T]$ and a.e. $x \in (0,\infty)$, and also for a.e. $(x,t) \in (0,\infty) \times (0,\infty)$.

By Corollary 6.6, $u_{m_k}(x,t)$ satisfies (33) and hence satisfies the integral equation obtained by integrating both sides of (33) over $s \in [0,t]$. As before, we take limits as $k \to \infty$ on both sides of this integrated version of (33) to obtain the corresponding integral equation satisfied by $u(x,t)$.

We choose $f_m = P_m f$ so that $f_{m_k}(x) \to f(x)$ for a.e. $x > 0$. Applying the above procedure we find that

$$\begin{aligned} u(x,t) = f(x) + \lim_{m_k \to \infty} \Big\{ &\int_0^t \int_x^\infty \gamma(y,x) u_{m_k}(y,\sigma) dy d\sigma - \int_0^t a(x) u_{m_k}(x,\sigma) d\sigma \\ &+ \frac{1}{2} \int_0^t \int_0^x K(x-y,y) u_{m_k}(x-y,\sigma) u_{m_k}(y,\sigma) dy d\sigma \\ &- \int_0^t u_{m_k}(x,\sigma) \int_0^{m_k - x} K(x,y) u_{m_k}(y,\sigma) dy d\sigma \Big\}. \end{aligned} \tag{39}$$

(i) Since $u_{m_k}(t) \to u(t)$ uniformly in t, $\int_0^t u_{m_k}(\sigma)\, d\sigma \to \int_0^t u(\sigma)\, d\sigma$ in X. Therefore, there exists a subsequence of $\{m_k\}$ which we also denote by $\{m_k\}$, such that

$$\left[\int_0^t u_{m_k}(\sigma)\, d\sigma\right](x) \to \left[\int_0^t u(\sigma)\, d\sigma\right](x) \quad \text{for a.e. } x > 0.$$

We deduce that $\int_0^t u_{m_k}(x,\sigma)\,d\sigma \to \int_0^t u(x,\sigma)\,d\sigma$ since X is a Banach space of type L, and so for a.e. $x \in (0,\infty)$, $\int_0^t a(x)u_{m_k}(x,\sigma)\,d\sigma \to \int_0^t a(x)u(x,\sigma)\,d\sigma$ as $m_k \to \infty$.

(ii) We can show that, for each fixed $\sigma \in [0,t]$, $g_{m_k}(\sigma) \to g(\sigma)$ strongly in X, where

$$[g_{m_k}(\sigma)](x) = \int_0^x K(x-y,y)u_{m_k}(x-y,\sigma)u_{m_k}(y,\sigma)dy,$$

and likewise for $[g(\sigma)](x)$, with u_{m_k} replaced by u. For each $m_k > 0$, $|||g_{m_k}(\sigma)||| \leq 2C_K b^2$. Thus, by the Lebesgue dominated convergence theorem for Bochner integrable functions [3, Thm. 3.7.9, p. 83], $\int_0^t g_{m_k}(\sigma)\,d\sigma \to \int_0^t g(\sigma)\,d\sigma$ in X. Again, there exists a subsequence of $\{m_k\}$ such that

$$\int_0^t [g_{m_k}(\sigma)](x)\,d\sigma \to \int_0^t [g(\sigma)](x)\,d\sigma.$$

Thus

$$\int_0^t\!\!\int_0^x K(x-y,y)u_{m_k}(x-y,\sigma)u_{m_k}(y,\sigma)\,dy\,d\sigma \to \int_0^t\!\!\int_0^x K(x-y,y)u(x-y,\sigma)u(y,\sigma)\,dy\,d\sigma.$$

(iii) In a similar manner to part (ii), we can show that

$$\int_0^t u_{m_k}(x,\sigma)\int_0^{m_k-x} K(x,y)u_{m_k}(y,\sigma)\,dy\,d\sigma \to \int_0^t u(x,\sigma)\int_0^\infty K(x,y)u(y,\sigma)\,dy\,d\sigma.$$

(iv) Substituting parts (i)–(iii) into (39), we know that

$$\lim_{m_k\to\infty}\int_0^t\int_x^\infty \gamma(y,x)u_{m_k}(y,\sigma)\,dy\,d\sigma$$

exists in $\mathbf{R}$. We can show that

$$\lim_{m_k\to\infty}\int_x^\infty \gamma(y,x)u_{m_k}(y,\sigma)\,dy = \int_x^\infty \gamma(y,x)u(y,\sigma)\ dy,$$

and that

$$\left|\int_x^\infty \gamma(y,x)u_{m_k}(y,\sigma)dy\right| \leq \tfrac{2}{3}C_F b.$$

Hence, by the Lebesgue dominated convergence theorem,

$$\int_0^t\int_x^\infty \gamma(y,x)u_{m_k}(y,\sigma)dyd\sigma \to \int_0^t\int_x^\infty \gamma(y,x)u(y,\sigma)dyd\sigma$$

as $m_k \to \infty$.

Substituting (i)–(iv) into (39), we complete the proof of Proposition 5.1. □

We conclude by commenting on the difficulty of proving the nonnegativity of the solution of the truncated system. If we attempt to follow the steps carried out in section 4, then we must first modify the definitions given in Notation 4.1. In particular, Notation 4.1(iii) is replaced by

$$\xi\colon (0,m)\times[0,T] \to \mathbf{R}$$

defined by $\xi(x,t) := -C_K \int_0^{m-x} [\tilde{u}_m(t)](y)\,dy$.

The problem we now face is that the function ξ depends on x as well as t. This means that

$$\big(S(t-\sigma)[\xi(\,.\,,\sigma)\tilde{u}_m(\sigma)]\big)(\,.\,) \neq \xi(\,.\,,\sigma)\big(S(t-\sigma)[\tilde{u}_m(\sigma)]\big)(\,.\,),$$

and so we cannot perform the step leading to (25). Had we been able to carry out this interchange, the result would have followed immediately from the results of Lemmas 4.2–4.8 and Theorem 4.9, after appropriate changes had also been made to the definitions of h, g, β, and α. It should, however, be noted that in their analysis of the truncated full equation for constant kernels, Aizenman and Bak [1] state that the solution to the truncated full equation can be shown to be nonnegative by the same argument as that used for the nontruncated equation; unfortunately, no details are supplied.

REFERENCES

[1] M. Aizenman and T.A. Bak, *Convergence to equilibrium in a system of reacting polymers*, Comm. Math. Phys., 65 (1979), pp. 203–230.

[2] J.A. Goldstein, *Semigroups of Linear Operators and Applications*, Oxford University Press, Oxford, 1985.

[3] E. Hille and R.S. Phillips, *Functional Analysis and Semigroups*, American Mathematical Society Colloquium Publications, AMS, Providence, RI, 1957.

[4] D.J. McLaughlin, *Coagulation and Fragmentation Models: A Semigroup Approach*, Ph.D. thesis, Univ. of Strathclyde, Glasgow, Scotland, February 1995.

[5] D.J. McLaughlin, W. Lamb, and A.C. McBride, *A semigroup approach to fragmentation models*, SIAM J. Math. Anal., 28 (1997), pp. 1158–1172.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1191–1204, September 1997

INVERSE BACKSCATTERING FOR THE ACOUSTIC EQUATION [*]

PLAMEN STEFANOV[†] AND GUNTHER UHLMANN[‡]

Abstract. In this paper, we consider the inverse backscattering problem for the acoustic equation. The problem is to determine the sound speed of a medium by measuring the response to sound waves in the backscattering direction, i.e., by measuring the echoes. We prove that one can uniquely identify the sound speed from this information if it is a priori close to the constant sound speed.

Key words. backscattering, inverse problems, acoustic equation

AMS subject classifications. 35R30, 25P25

PII. S0036141096301853

1. Introduction and statement of the results.

Consider the acoustic wave equation

$$(\partial_t^2 - c^2(x)\Delta)u = 0, \qquad (t,x) \in \mathbf{R}\times\mathbf{R}^3, \tag{1.1}$$

which describes the propagation of sound waves in an inhomogeneous medium with sound speed $c(x)$. We assume throughout the paper that $0 < c(x)$, $x \in \mathbf{R}^3$, and that for some $\rho > 0$ we have

$$c(x) = 1 \qquad \text{for } |x| \geq \rho. \tag{1.2}$$

The scattering kernel measures, roughly speaking, the effect of the inhomogeneity on an incident plane wave of the form $\delta(t - x\cdot\theta)$ with $\theta \in S^2$. More precisely, assume that $c \in C^2(\mathbf{R}^3)$ and let $u(t,x,\theta)$ be the solution of the Cauchy problem

$$\left\{\begin{array}{rcll} (\partial_t^2 - c^2(x)\Delta)u &=& 0, & (t,x)\in\mathbf{R}\times\mathbf{R}^3,\\ u|_{t\ll 0} &=& \delta(t - x\cdot\theta). & \end{array}\right. \tag{1.3}$$

We have that

$$u = \partial_t^3 w,$$

where $w(t,x,\theta)$ solves

$$\left\{\begin{array}{rcll} (\partial_t^2 - c^2(x)\Delta)w &=& 0, & (t,x)\in\mathbf{R}\times\mathbf{R}^3,\\ w|_{t\ll 0} &=& h_2(t - x\cdot\theta), & \end{array}\right.$$

with $h_2(s) = s^2/2$ for $s \geq 0$, and $h_2(s) = 0$ otherwise. We write

$$w = h_2(t - x\cdot\theta) + w_{\text{sc}}.$$

[*] Received by the editors March 22, 1996; accepted for publication June 5, 1996.
http://www.siam.org/journals/sima/28-5/30185.html

[†] Institute of Mathematics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria (stef@bgearn.acad.bg). This author was partly supported by the Bulgarian Research Foundation grant MM–407. The essential part of this paper was written while this author was visiting the University of Washington, Seattle.

[‡] Department of Mathematics, University of Washington, Seattle, WA 98195 (gunther@math.washington.edu). This author was partly supported by NSF grant DMS–9322619 and ONR grant N00014–93–1–0295.

In the Lax–Phillips theory of scattering [L-P] (see also [C-S], [P]) the *asymptotic wave profile* $w_{\mathrm{sc}}^{\#}$ of w_{sc} is defined by

$$w_{\mathrm{sc}}^{\#}(s,\omega,\theta) = \lim_{t\to\infty}(t+s)\partial_t w_{\mathrm{sc}}(t,(t+s)\omega,\theta),$$

where the limit exists in $L^2(\mathbf{R}_s \times S^2_\omega)$ for any $\theta \in S^2$. Then the *scattering kernel* is given by

$$S(s,\omega,\theta) = -\frac{1}{2\pi}\partial_s^3 w_{\mathrm{sc}}^{\#}(s,\omega,\theta).$$

We note that the scattering kernel S is closely connected with the Schwartz kernel of the scattering operator $\mathcal{S}$. In fact, $S(s'-s,\omega',\omega)$ is the Schwartz kernel of $\mathcal{R}(\mathcal{S}-I)\mathcal{R}^{-1}$, $\mathcal{R}$ being the Lax–Phillips translation representation [L-P] (see section 2).

The inverse backscattering problem consists in the determination of $c(x)$ from $S(s,-\theta,\theta)$, that is, roughly speaking, whether we can determine the sound speed by measuring the echoes produced by an incident plane wave in the direction θ. In this paper we show that measuring the echoes is enough to recover the sound speed if it is a priori close to a constant.

Theorem 1.1. *Let S_j be the scattering kernel associated with the sound speed c_j, $j=1,2$, satisfying* (1.2). *Assume further that $c_j \in W^{10,\infty}(\mathbf{R}^3)$. There exists $\varepsilon > 0$ such that if*

$$S_1(s,-\theta,\theta) = S_2(s,-\theta,\theta) \qquad \textit{for all } s\in\mathbf{R},\ \theta\in S^2,$$

and if

$$\|c_j - 1\|_{W^{10,\infty}(\mathbf{R}^3)} < \varepsilon, \qquad j=1,2,$$

then we have $c_1 = c_2$.

Guillemin proved in [G] that for the case considered here (and in more general situations) $\mathcal{S}$ is a Fourier integral operator, and he computed its symbol and canonical relation. In particular, $S(s,-\theta,\theta)$ makes sense and is a smooth function of θ with distributional values in the s-variable.

In the stationary approach to scattering, one considers the formal Fourier transform of (1.1):

$$\left(-\Delta + \lambda^2(1-c^{-2}(x)) - \lambda^2\right) v(x,\lambda) = 0. \tag{1.4}$$

Notice that one can consider (1.4) as a Schrödinger equation with potential

$$q(x) = \lambda^2(1-c^{-2}(x)).$$

However, this is not very useful for the study of the inverse backscattering problem, since we must consider high frequencies as well. The inverse scattering problem at a fixed energy has been solved in dimension $n \geq 3$ by Novikov [N]. This problem is in fact closely related to the inverse problem of determining a potential q from its associated Dirichlet to Neumann map. The latter problem was solved in [S-U]. For an account of this relationship, see, for instance, [U].

Given any $\theta\in S^2$, there are solutions of (1.4) of the form

$$v(x,\theta,\lambda) = e^{i\lambda x\cdot\theta} + \frac{e^{i\lambda|x|}}{|x|}a(\lambda,\omega,\theta) + o(|x|^{-1}), \qquad \text{as } |x|\to\infty, \tag{1.5}$$

where $\omega = x/|x|$. The function a is called the *scattering amplitude*. The relation between a and S is very simple:

$$\frac{i\lambda}{2\pi} a(\lambda, \omega, \theta) = \int e^{-is\lambda} S(s, \omega, \theta)\, ds.$$

Theorem 1.1 therefore has Theorem 1.2 as an immediate corollary.

THEOREM 1.2. *Let c_j, $j = 1, 2$, be as in Theorem* 1.1. *Let a_j denote the scattering amplitude associated with c_j, $j = 1, 2$. There exists $\varepsilon > 0$ such that if*

$$a_1(\lambda, -\theta, \theta) = a_2(\lambda, -\theta, \theta)$$

and if

$$\|c_j - 1\|_{W^{10,\infty}(\mathbf{R}^3)} < \varepsilon, \qquad j = 1, 2,$$

then $c_1 = c_2$.

The high-frequency asymptotics of the scattering amplitude have been considered in [G] and [V]. We do not know of any result for the inverse backscattering problem for the acoustic equation. The inverse backscattering problem for the Schrödinger equation has been studied in the papers [E-R], [B-L-M], [G-U], [M], [St II].

The structure of the paper is as follows. In section 2 we consider some preliminaries and prove Proposition 2.1, which gives a relation between $S_1 - S_2$ and $c_1^{-2} - c_2^{-2}$. In section 3 we construct the singular solution of (1.3). In section 4 we prove Theorem 1.1 by combining the results of section 3 and inverting a generalized Radon transform.

2. Preliminaries.

In this section we introduce the scattering kernel $S(s, \omega, \theta)$, and in Proposition 2.1, we prove a formula for the difference $S_1 - S_2$, where S_j, $j = 1, 2$, are related to two sound speeds $c_j \in C^2$ satisfying (1.2). A formula of a similar type related to a potential perturbation of the wave equation was first obtained in [St I].

The natural energy space for equation (1.1) is the completion $\mathcal{H}$ of $C_0^\infty(\mathbf{R}^3) \times C_0^\infty(\mathbf{R}^3)$ with respect to the energy norm

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2} \int \left(|\nabla f_1|^2 + c^{-2}(x) |f_2|^2 \right) dx, \qquad f = [f_1, f_2].$$

Throughout this paper we will denote two-dimensional vector functions ${}^t(f_1, f_2)$ by $[f_1, f_2]$. Then $\mathcal{H}$ is a Hilbert space and equation (1.1) is equivalent to

$$\partial_t u = -iAu, \qquad \text{with} \quad u = [u_1, u_2], \quad A = i \begin{pmatrix} 0 & I \\ c^2\Delta & 0 \end{pmatrix}; \tag{2.1}$$

i.e., if u solves (2.1), then $u_2 = \partial_t u_1$, $(\partial_t^2 - c^2\Delta)u_1 = 0$. Here I stands for the identity map. It is easy to see that A extends to a self-adjoint operator in $\mathcal{H}$; therefore, the solution to (2.1) is given by $u = e^{-itA} f =: U(t) f$, where $f = u|_{t=0}$. By Stone's theorem $U(t)$ forms a strongly continuous group of unitary operators in $\mathcal{H}$. Setting $c = 1$, we get the unperturbed group $U_0(t)$ in $\mathcal{H}_0$ related to the unperturbed wave equation $(\partial_t^2 - \Delta)u = 0$. The scattering operator $\mathcal{S}$ is then defined by $\mathcal{S} = W_-^{-1} W_+$, where the wave operators $W_\pm$ are defined as the strong limits $W_\pm = \text{s-lim}_{t \to \pm} U(t) U_0(-t)$. It is well known that the wave operators exist as bounded operators and moreover, $\mathcal{S}$ is also well defined as a bounded operator in $\mathcal{H}_0$ [L-P], [R-S].

As in the Introduction, we consider the scattering solution $u(t,x,\theta)$ as the solution to the following Cauchy problem:

$$\left\{\begin{array}{rcll} (\partial_t^2 - c^2\Delta)u &=& 0, & \text{in } \mathbf{R}_t \times \mathbf{R}_x^3,\\ u|_{t\ll 0} &=& \delta(t - x\cdot\theta). & \end{array}\right. \tag{2.2}$$

Here $\theta \in S^2$ is a parameter giving the direction of the incident plane wave in (2.2). The initial condition above can be replaced by $u|_{t=-\rho} = \delta(-\rho - x\cdot\theta)$, $u_t|_{t=-\rho} = \delta'(-\rho - x\cdot\theta)$. The standard way of constructing a solution of (2.2) is the following. Set $h_j(t) = t^j/j!$ for $t \ge 0$, and $h_j(t) = 0$ otherwise. Then $h_j' = h_{j-1}$, $j \ge 1$, and h_0 is the Heaviside function. If we replace the Dirac delta function δ in (2.2) by h_2, we get initial data $[h_2(-\rho - x\cdot\theta), h_1(-\rho - x\cdot\theta)]$ for $t = -\rho$, that belong locally to $\mathcal{H}$ and even to $D(A)$. As in the Introduction, consider the problem

$$\left\{\begin{array}{rcll} (\partial_t^2 - c^2\Delta)w &=& 0 & \text{in } \mathbf{R}_t \times \mathbf{R}_x^3,\\ w|_{t\ll 0} &=& h_2(t - x\cdot\theta). & \end{array}\right. \tag{2.3}$$

Then $w = h_2(t - x\cdot\theta) + w_{\text{sc}}$, where $(\partial_t^2 - c^2\Delta)w_{\text{sc}} = -(1-c^2)h_0(t - x\cdot\theta)$ and $w_{\text{sc}}|_{t\ll 0} = 0$. Therefore,

$$[w_{\text{sc}}, \partial_t w_{\text{sc}}] = -\int_{-\infty}^{t} U(t-s)(1-c^2)[0, h_0(s - x\cdot\theta)]\,ds. \tag{2.4}$$

Here $1-c^2$ has compact support; thus $(1-c^2)[0, h_0(s-x\cdot\theta)] \in \mathcal{H}$. Having constructed a solution to (2.3) we can now solve (2.2) by setting

$$u(t,x,\theta) = \partial_t^3 w(t,x,\theta). \tag{2.5}$$

Following Lax and Phillips [L-P] (see also [C-S]), as in the Introduction, we define the *asymptotic wave profile* $w_{\text{sc}}^{\#}$ of w_{sc} by

$$w_{\text{sc}}^{\#}(s,\omega,\theta) = \lim_{t\to\infty}(t+s)\partial_t w_{\text{sc}}(t,(t+s)\omega,\theta). \tag{2.6}$$

The limit exists in $L^2(\mathbf{R}_s \times S_\omega^2)$ for any θ [L-P], [C-S]. Then we define the scattering kernel S by

$$S(s,\omega,\theta) = -\frac{1}{2\pi}\partial_s^3 w_{\text{sc}}^{\#}(s,\omega,\theta). \tag{2.7}$$

In some sense S satisfies the asymptotics

$$\partial_t u(t,x,\theta) = \delta'(t - x\cdot\theta) - \frac{2\pi}{|x|}S\left(|x| - t, \frac{x}{|x|}, \theta\right) + o\left(\frac{1}{|x|}\right), \qquad \text{as } t, |x| \to \infty.$$

The formula above is a time-dependent analogue of the definition (1.5) of the scattering amplitude via the asymptotics of the solution v of the Lipmann–Schwinger equation for large x.

It turns out that S is closely related to the distribution kernel of the scattering operator $\mathcal{S}$. Denote by $(Rf)(s,\omega) = \int f(x)\delta(s - x\cdot\omega)dx$ the Radon transform of f and consider the operator $\mathcal{R}$ (the Lax and Phillips translation representation) defined by $\mathcal{R}[f_1,f_2] = \frac{1}{4\pi}(-\partial_s^2 Rf_1 + \partial_s Rf_2)$. Then $\mathcal{R}$ is a unitary map $\mathcal{R} : \mathcal{H}_0 \to L^2(\mathbf{R}\times S^2)$.

A well-known fact from the Lax–Phillips theory is that $S(s'-s,w',w)$ is the Schwartz kernel of $\mathcal{R}(\mathcal{S}-I)\mathcal{R}^{-1}$ (see [L-P], [C-S], [P]); i.e., in a distribution sense, we have

$$\left(\mathcal{R}(\mathcal{S}-I)\mathcal{R}^{-1}k\right)(s',\omega') = \int_{\mathbf{R}\times S^2} S(s'-s,\omega',\omega)k(s,\omega)\,ds\,d\omega. \tag{2.8}$$

Next we will derive a formula for $S_1 - S_2$, where S_j is related to c_j, $j=1,2$. Let us first notice that $(2\pi)^{-1}[u(\pm t\pm s,x,\pm\theta),\partial_t u(\pm t\pm s,x,\pm\theta)]$ is the distribution kernel of $U(t)W_\pm\mathcal{R}^{-1}$; i.e., for any $k\in C_0^\infty(\mathbf{R}\times S^2)$ in a distribution sense we have

$$U(t)W_\pm\mathcal{R}^{-1}k = \frac{1}{2\pi}\int_{\mathbf{R}\times S^2}\Big[u(\pm t\pm s,x,\pm\theta),\partial_t u(\pm t\pm s,x,\pm\theta)\Big]k(s,\theta)\,ds\,d\theta. \tag{2.9}$$

Indeed, denote $f=\mathcal{R}^{-1}k$ and consider W_+. Then $U(t)W_+\mathcal{R}^{-1}k = U(t+T)U_0(-T)f$ for some fixed $T>0$ depending on $\operatorname{supp} k$. Denote $[v,\partial_t v] = U(t+T)U_0(-T)f$ and denote also the right-hand side of (2.9) by $[\tilde v,\partial_t\tilde v]$. Both v and $\tilde v$ solve (1.1). Next, for $t<-T$, we have $[v,\partial_t v]=U_0(t)f$. On the other hand, for $t\ll 0$ we get for $\tilde v$

$$[\tilde v,\partial_t\tilde v] = \frac{1}{2\pi}\int_{\mathbf{R}\times S^2}\Big[\delta(t+s-x\cdot\theta),\delta'(t+s-x\cdot\theta)\Big]k(s,\theta)\,ds\,d\theta = U_0(t)f$$

by the inversion formula for $\mathcal{R}$ (see [L-P]). Therefore, v and $\tilde v$ have the same initial data and must coincide. This proves (2.9) for W_+. The proof for W_- is similar.

PROPOSITION 2.1. *Let $S_j(s,\omega,\theta)$ be the scattering kernel related to $c_j(x)\in C^2$, $j=1,2$. Then*

$$(S_1-S_2)(s,\omega,\theta) = \frac{1}{8\pi^2}\partial_s^3\iint(c_1^{-2}-c_2^{-2})u_1(t,x,\theta)u_2(-s-t,x,-\omega)\,dt\,dx,$$

where u_j are the scattering solutions related to c_j, $j=1,2$, and the integral is to be considered in a distribution sense.

Proof. Denote by $U_j(t)$, $j=1,2$, the propagators related to c_j. Consider the function $F(t)=U_2(T+t)U_1(-t+T)f$, $f\in D(A_1)=D(A_2)$. Then $F'(t)=-iU_2(T+t)(A_2-A_1)U_1(-t+T)$, and from $F(T)-F(-T)=\int_{-T}^T F'(t)dt$, we get

$$(U_2(2T)-U_1(2T))f = \int_{-T}^{T}U_2(T+t)QU_1(-t+T)f\,dt, \tag{2.10}$$

where

$$Q = \begin{pmatrix} 0 & 0 \\ (c_2^2-c_1^2)\Delta & 0\end{pmatrix}.$$

Next, choose two functions $k,l\in C_0^\infty(\mathbf{R}\times S^2)$ and set $f=\mathcal{R}^{-1}k$, $g=\mathcal{R}^{-1}l$. Then, by using standard arguments from the Lax–Phillips theory, we get that

$$(\mathcal{S}_j f,g)_{\mathcal{H}_0} = \Big(U_0(-T)U_j(2T)U_0(-T)f,g\Big)_{\mathcal{H}_0}$$

with some large $T>0$ depending on $\operatorname{supp} k$, $\operatorname{supp} l$. Therefore, by (2.10),

$$\begin{aligned}((\mathcal{S}_2-\mathcal{S}_1)f,g)_{\mathcal{H}_0} &= \int_{-T}^{T}\Big(U_0(-T)U_2(T+t)QU_1(-t+T)U_0(-T)f,g\Big)_{\mathcal{H}_0}dt\\ &= \int_{-T}^{T}\Big(QU_1(-t+T)U_0(-T)f,U_2(-t-T)U_0(T)g\Big)_{\mathcal{H}_2}dt.\end{aligned} \tag{2.11}$$

Here $\mathcal{H}_j$, $j = 0, 1, 2$, are related to $c_0 = 1$, c_1, and c_2, respectively. Next, note that $U_1(-t+T)U_0(-T)f = U_1(-t)W_+^{(1)}f = U_1(-t)W_+^{(1)}\mathcal{R}^{-1}k$. Similarly, $U_2(-t-T)U_0(T)g = U_2(-t)W_-^{(2)}\mathcal{R}^{-1}l$. Using (2.9), we get from (2.11)

$$\begin{aligned}((\mathcal{S}_2 - \mathcal{S}_1)f, g)_{\mathcal{H}_0} &\\ &= \frac{1}{8\pi^2}\int_{-T}^{T}\int\dots\int (c_2^2 - c_1^2)(\Delta u_1)(-t+s_1, x, \theta_1)\partial_t u_2(t-s_2, x, -\theta_2)\\ &\qquad \times k(s_1,\theta_1)l(s_2,\theta_2)c_2^{-2}\, ds_1 d\theta_1 ds_2 d\theta_2 dx\, dt\\ &= \frac{1}{8\pi^2}\int_{-T}^{T}\int\dots\int (c_1^{-2} - c_2^{-2})\partial_{s_1}^2 u_1(-t+s_1, x, \theta_1)\partial_t u_2(t-s_2, x, -\theta_2)\\ &\qquad \times k(s_1,\theta_1)l(s_2,\theta_2)\, ds_1 d\theta_1 ds_2 d\theta_2 dx\, dt. \end{aligned} \tag{2.12}$$

Clearly, the integrand above vanishes for $|t| > T$, so we may integrate in t over the whole real line. According to (2.8),

$$\begin{aligned}((\mathcal{S}_2 - \mathcal{S}_1)f, g)_{\mathcal{H}_0} &\\ &= \int_{[\mathbf{R}\times S^2]^2} (S_2 - S_1)(s_2 - s_1, \theta_2, \theta_1)k(s_1,\theta_1)l(s_2,\theta_2)\, ds_1 d\theta_1 ds_2 d\theta_2.\end{aligned} \tag{2.13}$$

Comparing (2.12) and (2.13), we conclude that

$$\begin{aligned}(S_1 - S_2)(s_2 - s_1, \theta_2, \theta_1) &\\ &= \frac{1}{8\pi^2}\iint (c_1^{-2} - c_2^{-2})\partial_{s_1}^2 u_1(-t+s_1, x, \theta_1)\partial_t u_2(t-s_2, x, -\theta_2)\, dx\, dt.\end{aligned}$$

The right-hand side above as a function of s_1, s_2 depends merely on $s_2 - s_1$, and setting $s = s_2 - s_1$, $\tilde{t} = -t + s_1$, we complete the proof of the proposition. □

3. Singular decomposition of the scattering solution. In this section we prove that the scattering solution $u(t, x, \theta)$ admits a singular decomposition of the type $u(t, x, \theta) = \alpha(x, \theta)\delta(t - \phi(x, \theta)) + \beta(x, \theta)h_0(t - \phi(x, \theta)) + r(t, x, \theta)$, where ϕ is a suitable phase function and the remainder $r(t, \cdot, \theta)$ belongs to $H^1 \cap L^\infty$, $\partial_t r \in L^2$. Such decompositions are in principle known for that kind of problem (see, e.g., [V] for a high-frequency asymptotics of the solution v of (1.4) given by (1.5)). Our goal here is to prove estimates on the remainder which are uniform in $c(x)$ under the assumption of a finite smoothness of c. As in Theorem 1.1, we assume that c is close to $c = 1$ in the $W^{m,\infty}$ topology for some m. It turns out that in our proof we need estimates on the remainder for t belonging to a finite interval only. This fact considerably simplifies our analysis. On the other hand, in principle one could obtain estimates on the remainder for large t which are also uniform in c. This is related to the problem of finding estimates of the remainder in the high-frequency asymptotics of the solution v of (1.4) defined in (1.5) (see [V]) which are uniform in c or finding estimates on the resolvent of $c^2\Delta + \lambda^2$. The latter problems are more delicate ones. In fact, one of the main reasons for working with time-dependent methods is the advantage we get by dealing with bounded t's only.

We start with analysis of the phase function ϕ related to (1.1). We define $\phi(x, \theta)$ as the solution to the eikonal equation

$$\left\{\begin{array}{rcl} (\nabla\phi)^2 &=& c^{-2}(x),\\ \phi|_{x\cdot\theta\ll 0} &=& x\cdot\theta. \end{array}\right. \tag{3.1}$$

Throughout this section we assume that c satisfies (1.2) and that

$$\|c-1\|_{W^{m+1,\infty}} < \varepsilon \tag{3.2}$$

with some $\varepsilon > 0$ and $m \geq 2$. We need to solve (3.1) in B_ρ. Fix $\theta \in S^2$. We may assume that $\theta = {}^t(1,0,0)$. Then (3.1) can be rewritten as

$$\left\{\begin{array}{rcl} (\nabla\phi)^2 &=& c^{-2}(x), \\ \phi|_{x_1=-\rho} &=& -\rho, \\ \partial_{x_1}\phi|_{x_1=-\rho} &=& 1. \end{array}\right. \tag{3.3}$$

The Hamiltonian system associated with (3.3) is

$$\left\{\begin{array}{rclrcll} \frac{d}{ds}x &=& 2\xi, & \frac{d}{ds}\xi &=& \nabla c^{-2}, & \\ x|_{s=0} &=& {}^t(-\rho,\eta), & \xi|_{s=0} &=& {}^t(1,0,0), & \eta \in \mathbf{R}^2. \end{array}\right. \tag{3.4}$$

Notice that the solution to (3.4) in the case $c = 1$ is $x = {}^t(2s-\rho,\eta)$, $\xi = {}^t(1,0,0)$. On the other hand, for general $c(x)$, the solution of (3.4) exists for any s (see [V]).

LEMMA 3.1. *Fix $a > 0$. Then there exists $C > 0$ such that for the solution $x = x(s,\eta)$, $\xi = \xi(s,\eta)$ of* (3.4) *we have*

$$\|x - {}^t(2s-\rho,\eta)\|_{W^{m,\infty}([0,a]\times\mathbf{R}^2)} + \|\xi - {}^t(1,0,0)\|_{W^{m,\infty}([0,a]\times\mathbf{R}^2)} \leq C\varepsilon.$$

The proof of the lemma is based on a comparison theorem for ODEs and will be omitted here.

In particular, Lemma 3.1 implies that under the smallness assumption (3.2) the Hamiltonian flow is nontrapping for small ε, more precisely, $x(s,\eta) \notin B_\rho = \{x;\ |x| < \rho\}$ for $s > a$ with some $a > 0$. Moreover, the mapping ${}^t(s,\eta) \mapsto x(s,\eta)$ is a $W^{m,\infty}$-diffeomorphism on $[0,a]\times\{\eta \in \mathbf{R}^2;\ |\eta| \leq 2\rho\}$, and its range covers B_ρ, provided that ε is small enough. We will need, in fact, to work in a larger domain, so let us assume that ε and a are such that ${}^t(s,\eta) \mapsto x(s,\eta)$ maps $[0,a]\times\{\eta \in \mathbf{R}^2;\ |\eta| \leq 5\rho\}$ into a compact covering $B_{4\rho}$. The phase function ϕ solving (3.3) is defined in $B_{4\rho}$ by (see [V])

$$\phi = -\rho + 2\int c^{-2}(x)\,ds,$$

where the integration is taken over the shortest characteristics $x = x(s,\eta)$ joining the plane $x_1 = -\rho$ and x. The change of coordinates $x \mapsto {}^t(s,\eta)$ is ε-close to $x = {}^t(2s-\rho,\eta)$ in $W^{m,\infty}$, which easily implies that ϕ must be close to $\phi = x_1$. So far θ was fixed. One can also examine easily the dependence of ϕ on $\theta \in S^2$. Thus we get the following lemma.

LEMMA 3.2. *Assume that* (3.2) *holds with $\varepsilon > 0$ sufficiently small. Then there exists $C_0 > 0$ such that*

$$\|\phi(x,\theta) - x\cdot\theta\|_{W^{m,\infty}(B_{4\rho}\times S^2)} \leq C_0\varepsilon.$$

Now we are ready to prove the principal result of this section about the scattering solution $u(t,x,\theta)$ introduced in (2.2). Denote

$$T = \rho + C_0\varepsilon, \tag{3.5}$$

where C_0 is the constant in Lemma 3.2. Note that $\max\{|\phi(x,\theta)|;\ x\in B_\rho,\ \theta\in S^2\}\le T$.

PROPOSITION 3.3. *Assume that* (3.2) *holds with $m\ge 9$ and $\varepsilon>0$ sufficiently small. Then there exists a constant $C>0$, such that for $|t|<3T$, and for any $\theta\in S^2$, we have*

$$u(t,x,\theta)=\alpha(x,\theta)\delta(t-\phi(x,\theta))+\beta(x,\theta)h_0(t-\phi(x,\theta))+r(t,x,\theta),$$

where

$$\|\alpha-1\|_{W^{m-2,\infty}(B_{4\rho}\times S^2)}\le C\varepsilon,\qquad |\beta(x,\theta)|\le C\varepsilon, \tag{3.6}$$

and

$$\|r(t,\cdot,\theta)\|_{L^\infty}+\|\partial_t r(t,\cdot,\theta)\|_{L^2}\le C\varepsilon. \tag{3.7}$$

Proof. Let us look for u of the form

$$u(t,x,\theta)=\alpha(x,\theta)\delta(t-\phi(x,\theta))+\beta(x,\theta)h_0(t-\phi(x,\theta))+\gamma(x,\theta)h_1(t-\phi(x,\theta))+\tilde r(t,x,\theta).$$

Then $\alpha=1+\tilde\alpha$, β, γ solve the transport equations

$$(2\nabla\phi\cdot\nabla+\Delta\phi)\tilde\alpha=-\Delta\phi,\qquad \tilde\alpha|_{x\cdot\theta=-\rho}=0, \tag{3.8}$$

$$(2\nabla\phi\cdot\nabla+\Delta\phi)\beta=\Delta\alpha,\qquad \beta|_{x\cdot\theta=-\rho}=0, \tag{3.9}$$

$$(2\nabla\phi\cdot\nabla+\Delta\phi)\gamma=\Delta\beta,\qquad \gamma|_{x\cdot\theta=-\rho}=0, \tag{3.10}$$

while $\tilde r$ solves

$$(c^{-2}\partial_t^2-\Delta)\tilde r=(\Delta\gamma)h_1(t-\phi),\qquad \tilde r|_{t\ll 0}=0. \tag{3.11}$$

Note that we need to solve (3.8)–(3.10) in the compact $x\cdot\theta\ge-\rho$, $\phi(x,\theta)\le 3T$, $|\eta|<\rho$ ($\eta=\eta(x)$ is determined by $x=x(s,\eta)$), and for ε sufficiently small this compact is contained in $B_{4\rho}$, where ϕ is well defined. The first equation (3.8) can be solved in $B_{4\rho}$ and (3.6) follows directly from Lemma 3.2. The estimate (3.6) for α follows easily from Lemmas 3.1 and 3.2. Next, since $\Delta\alpha=O(\varepsilon)$, we get (if $m\ge 4$) (3.6) for β as well. Similarly, if $m\ge 6$, then $|\gamma|=O(\varepsilon)$ as well. Finally, for $\tilde r$ we get by (3.11)

$$[\tilde r,\partial_t\tilde r]=\int_{-\rho}^t U(t-s)[0,(\Delta\gamma)h_1(s-\phi)]\,ds.$$

We get as above that $(\Delta\gamma)h_1(s-\phi)$ is supported in $B_{4\rho}$ for $-\rho\le s\le t$, $|t|<3T$, and moreover $\|[0,(\Delta\gamma)h_1(s-\phi)]\|_{\mathcal H}\le C\varepsilon$ (if $m\ge 8$). Note that the norm in $\mathcal H$ depends on $c(x)$ but is uniformly bounded when c satisfies (3.2) with $\varepsilon<1$. So we get

$$\|[\tilde r,\partial_t\tilde r]\|_{\mathcal H}\le C(t+\rho)\varepsilon,\qquad -\rho\le t\le T \tag{3.12}$$

(and $\tilde{r} = 0$ for $t < -\rho$). Next, $[\tilde{r}, \partial_t\tilde{r}] \in D(A)$ and

$$A[\tilde{r}, \partial_t\tilde{r}] = [\partial_t\tilde{r}, c^2\Delta\tilde{r}] = \int_{-\rho}^{t} U(t-s)A[0, (\Delta\gamma)h_1(s-\phi)]\, ds$$
$$= \int_{-\rho}^{t} U(t-s)[(\Delta\gamma)h_1(s-\phi), 0]\, ds.$$

Since $\|[(\Delta\gamma)h_1(s-\phi), 0]\|_{\mathcal{H}} = O(\varepsilon)$ (here we need $m = 9$), we get as above that

$$\big\|[\partial_t\tilde{r}, c^2\Delta\tilde{r}]\big\|_{\mathcal{H}} \le C(t+\rho)\varepsilon, \qquad -\rho \le t \le T. \tag{3.13}$$

By (3.12) and (3.13),

$$\|\nabla\tilde{r}\| + \|\Delta\tilde{r}\| + \|\partial_t\tilde{r}\| + \|\nabla\partial_t\tilde{r}\| \le C\varepsilon,$$

where $\|\cdot\| = \|\cdot\|_{L^2}$. Moreover, $\tilde{r}$ is compactly supported (uniformly in $\varepsilon < 1$, $|t| < 3T$) because of the finite speed of propagation for (1.1). Therefore, by the Poincaré inequality (see, e.g., [L-P]), we get $\|\tilde{r}\| = O(\varepsilon)$ as well. Thus,

$$\|\tilde{r}\|_{H^2} + \|\partial_t\tilde{r}\|_{H^1} \le C\varepsilon.$$

By the Sobolev embedding theorem this yields $\|\tilde{r}\|_{L^\infty} + \|\partial_t\tilde{r}\|_{L^2} = O(\varepsilon)$, and combining this with (3.6), we get (3.7) for $r = \gamma h_1(t-\phi) + \tilde{r}$. □

4. Proof of Theorem 1.1. Assume that the hypotheses of Theorem 1.1 are fulfilled and denote by u_j the scattering solutions related to c_j, $j = 1, 2$. Then, by Proposition 2.1,

$$\iint q(x)u_1(t,x,\theta)u_2(s-t,x,\theta)\, dx\, dt = 0, \qquad q := c_1^{-2} - c_2^{-2} \tag{4.1}$$

for any $s \in \mathbf{R}$, $\theta \in S^2$. Let us now apply Proposition 3.3 and substitute u_j, $j = 1, 2$, in (4.1) by its singular expansion. We get

$$\begin{aligned} -\int q\alpha_1\alpha_2\delta(s-\phi_1-\phi_2)\, dx & \\ = \int q\Big[&\alpha_2\beta_1 h_0(s-\phi_1-\phi_2) + \alpha_1\beta_2 h_0(s-\phi_1-\phi_2) \\ &+ \alpha_2 r_1(s-\phi_2) + \alpha_1 r_2(s-\phi_1)\Big]\, dx \\ + \iint q\Big[&\beta_1\beta_2 h_0(t-\phi_1)h_0(s-t-\phi_2) + r_1(t)r_2(s-t) \\ &+ \beta_1 h_0(t-\phi_1)r_2(s-t) + \beta_2 h_0(s-t-\phi_2)r_1(t)\Big]\, dx\, dt. \end{aligned} \tag{4.2}$$

Here $r_1(t) = r_1(t,x,\theta)$, $\phi_1 = \phi_1(x,\theta)$, etc. Denote $\phi(x,\theta) = \phi_1(x,\theta) + \phi_2(x,\theta)$, $a(x,\theta) = \alpha_1(x,\theta) + \alpha_2(x,\theta)$. Since by Lemma 3.2, $\phi(x,\theta)$ is close to $2x\cdot\theta$ and $a(x,\theta)$ is close to 1, the left-hand side of (4.2) reminds us of the Radon transform Rq of q. Let us recall that we have the following Parseval's equality for the Radon transform $\|\partial_s Rf\|_{L^2(\mathbf{R}\times S^2)} = 4\pi\|f\|_{L^2}$. Bearing this in mind, let us differentiate (4.2) with respect to s.

$$-\partial_s \int qa\delta(s-\phi)\, dx = I_1 + I_2 + I_3 + I_4, \tag{4.3}$$

where

$$\begin{aligned} I_1 &= \int q(\alpha_2\beta_1 + \alpha_1\beta_2)\delta(s-\phi)\,dx, \\ I_2 &= \int q[\alpha_2\partial_s r_1(s-\phi_2) + \alpha_1\partial_s r_2(s-\phi_1)]\,dx, \\ I_3 &= \int q[\beta_1\beta_2 h_0(s-\phi) + \beta_1 r_2(s-\phi_1) + \beta_2 r_1(s-\phi_2)]\,dx, \\ I_4 &= \iint q r_1(t)\partial_s r_2(s-t)\,dx\,dt. \end{aligned}$$

The left-hand side of (4.3) vanishes for $|s| > 2T$ (see Lemma 3.2 and (3.5)). Therefore, so does the right-hand side above, but this is not necessarily true for each term I_j. Let us estimate the norm in $L^2([-2T,2T]\times S^2)$ of each term in (4.3). For the left-hand side in (4.3) we have

$$\begin{aligned} &\left\| \partial_s \int q(x)a(x,\theta)\delta(s-\phi(x,\theta))\,dx \right\|_{L^2([-2T,2T]\times S^2)} \\ &\qquad = (2\pi)^{-1/2} \left\| k \int e^{ik\phi(x,\theta)} a(x,\theta) q(x)\,dx \right\|_{L^2(\mathbf{R}_k \times S^2_\theta)}. \end{aligned} \tag{4.4}$$

Let us extend $\phi(x,\xi)$ and $a(x,\theta)$ for $\xi \notin S^2$ by $\phi(x,\xi) = |\xi|\phi(x,\xi/|\xi|)$ and $a(x,\xi) = a(x,\xi/|\xi|)$, respectively. Then Lemma 3.2 implies

$$\left| \partial_x^\alpha \partial_\xi^\beta (\phi(x,\xi) - 2x\cdot\xi) \right| \le C_1 \varepsilon |\xi|^{1-|\beta|} \quad \text{for } |\alpha| + |\beta| \le m,\ x \in B_{4\rho},\ \xi \neq 0. \tag{4.5}$$

Similarly, (3.6) implies

$$\left| \partial_x^\alpha \partial_\xi^\beta (a(x,\xi) - 1) \right| \le C_1 \varepsilon |\xi|^{-|\beta|} \quad \text{for } |\alpha| + |\beta| \le m-2,\ x \in B_{4\rho},\ \xi \neq 0. \tag{4.6}$$

Since q is real-valued, the square integral of the expression in the right-hand side of (4.4) over $\mathbf{R}_k \times S^2$ equals twice the square integral over $\mathbf{R}_k^+ \times S^2_\theta$. Setting $\xi = k\theta$, $k > 0$, $\theta \in S^2$, we obtain from (4.4)

$$\left\| \partial_s \int q(x)a(x,\theta)\delta(s-\phi(x,\theta))\,dx \right\|_{L^2([-2T,2T]\times S^2)} = \sqrt{2}(2\pi)^{-1/2} \|Pq\|_{L^2(\mathbf{R}^3_\xi)}, \tag{4.7}$$

where

$$(Pq)(\xi) = \int e^{i\phi(x,\xi)} a(x,\xi) q(x)\,dx. \tag{4.8}$$

Our plan is the following. First we will show that $C_1\|q\| \le \|Pq\| \le C_2\|q\|$ with some $C_1 > 0$, $C_2 > 0$ independent of ε. Next we are going to estimate the norms in $L^2([-2T,2T]\times S^2)$ of each term $I_j = I_j(s,\theta)$ in (4.3) and will show that $I_j = O(\varepsilon\|q\|)$, $j = 1,2,3,4$. Then (4.3), (4.7) would imply that $C_1\|q\| \le \|Pq\| \le C\varepsilon\|q\|$; hence $q = 0$.

Proposition 4.1. *If c_j, $j = 1,2$, satisfy (3.2) with $m = 9$ and if $\varepsilon > 0$ is sufficiently small, then $P : L^2(B_\rho) \to L^2(\mathbf{R}^3_\xi)$ is a bounded operator. Moreover, there exist two constants $C_1 > 0$, $C_2 > 0$ independent of ε (small enough), c_1, c_2, such that*

$$C_1\|f\| \le \|Pf\| \le C_2\|f\| \qquad \textit{for any } f \in L^2(B_\rho).$$

Proof. We will show that the estimate above follows from the fact that $\phi = \phi_1 + \phi_2$ is close to $2x \cdot \theta$ (see Lemma 3.2) and a is close to 1 (see 4.6). This does not necessarily imply that P (see (4.8)) is close to the Fourier transform, but one can expect that P^*P is close to cI with some constant c. We have

$$(4.9) \qquad (P^*Pf)(x) = \iint e^{-i(\phi(x,\xi)-\phi(y,\xi))} a(x,\xi)a(y,\xi)f(y)\,dy\,d\xi.$$

The phase function above admits the representation

$$\phi(x,\xi) - \phi(y,\xi) = 2(x-y)\cdot\eta(x,y,\xi),$$

where

$$(4.10) \qquad \eta(x,y,\xi) = \frac{1}{2}\int_0^1 (\nabla_x\phi)(y+t(x-y),\xi)\,dt.$$

To prove (4.10), it is enough to apply the identity $g(1) - g(0) = \int_0^1 g'(t)dt$ to the function $g(t) = \phi(y + t(x-y))$. By Lemma 3.2, $\eta(x,y,\xi)$ belongs to $W^{m-1,\infty}$ and is homogeneous with respect to ξ of order one. Moreover,

$$\left|\partial_x^\alpha \partial_y^\beta \partial_\xi^\gamma (\eta(x,y,\xi) - \xi)\right| \le C\varepsilon|\xi|^{1-|\gamma|}$$

for $|\alpha| + |\beta| + |\gamma| \le m-1$, $x \in B_{4\rho}$, $y \in B_{4\rho}$, $\xi \neq 0$. The equation $\eta = \eta(x,y,\xi)$ can be solved for ξ provided that ε is sufficiently small. The Jacobian $J := |D\eta/D\xi|$ satisfies the estimates

$$(4.11) \qquad \left|\partial_x^\alpha \partial_y^\beta \partial_\xi^\gamma (J(x,y,\xi) - 1)\right| \le C\varepsilon|\xi|^{-|\gamma|}$$

for $|\alpha| + |\beta| + |\gamma| \le m-2$, $x \in B_{4\rho}$, $y \in B_{4\rho}$, $\xi \neq 0$. Let us perform the change of variables $\xi \to \eta$ in (4.9):

$$(4.12) \qquad P^*Pf = \iint e^{-2i(x-y)\cdot\eta} b(x,y,\eta)f(y)\tilde{J}(x,y,\eta)\,dy\,d\eta,$$

where $\tilde{J}(x,y,\eta) = J^{-1}(x,y,\xi)|_{\xi=\xi(x,y,\eta)}$, $b(x,y,\eta) = a(x,\xi)a(y,\xi)|_{\xi=\xi(x,y,\eta)}$. The principal part of the integral above is

$$\iint e^{-2i(x-y)\cdot\eta} f(y)\,dy\,d\eta = \pi^3 f,$$

so from (4.12) we get

$$(4.13) \qquad \left(P^*P - \pi^3 I\right) f = \iint e^{-2i(x-y)\cdot\eta} f(y)\left((b\tilde{J})(x,y,\eta) - 1\right)\,dy\,d\eta.$$

We are going to apply Theorem A.1 (see the Appendix below) to (4.13). By (4.11), (3.6),

(4.14)
$$\left|\partial_x^\alpha \partial_y^\beta \left((b\tilde{J})(x,y,\eta) - 1\right)\right| \le C\varepsilon \quad \text{for } |\alpha| + |\beta| \le m-2,\ x \in B_{4\rho},\ y \in B_{4\rho},\ \eta \neq 0.$$

Let us extend the operator $P^*P-\pi^3I$, defined a priori on $L^2(B_\rho)$ to an operator Q in $L^2(\mathbf{R}^3)$ by (4.13) with $\tilde{J}-1$ replaced by $\chi(x)(\tilde{J}-1)\chi(y)$, where $\chi\in C_0^\infty$, $\operatorname{supp}\chi\subset B_{2\rho}$, $\chi=1$ on B_ρ. Then, if $m-2=7$, Theorem A.1 yields $\|Q\|_{\mathcal{L}(L^2(\mathbf{R}^3))}\le C\varepsilon$, which implies

$$\|P^*P-\pi^3I\|_{\mathcal{L}(L^2(B_\rho))}\le C\varepsilon.$$

Thus, for any $f\in L^2(B_\rho)$, we have

$$\big|\|Pf\|^2-\pi^3\|f\|^2\big|=\big|\big(P^*Pf-\pi^3f,f\big)\big|\le C\varepsilon\|f\|^2,$$

and this completes the proof of Proposition 4.1 for ε small enough. □

We proceed now with estimating the norms of I_j, $j=1,2,3,4$, in $L^2([-2T,2T]\times S^2)$. By (3.6) and (4.7) we get for I_1

$$\begin{aligned}\|I_1\|_{L^2([-2T,2T]\times S^2)}&\le C\varepsilon\left\|\int|q|\delta(s-\phi)\,dx\right\|_{L^2(\mathbf{R}\times S^2)}\\&\le C'\varepsilon\left\|\partial_s\int|q|\delta(s-\phi)\,dx\right\|_{L^2(\mathbf{R}\times S^2)}\\&\le C''\|P_0|q|\|\le C'''\|q\|.\end{aligned}\tag{4.15}$$

Here P_0 is the operator (4.8) with $a=1$. In order to prove (4.15), we have approximated $|q|$ with smooth functions and have used the fact that for any $f\in C^1(\mathbf{R})$ with $f=0$ outside some finite interval $[-a,a]$, we have $\|f\|_{L^2}\le C(a)\|f'\|_{L^2}$.

To estimate I_2, I_3, and I_4, observe that

$$I_2+I_3+I_4=\int K(s,\theta,x)q(x)\,dx\tag{4.16}$$

with

$$\begin{aligned}K&=\alpha_2\partial_sr_1(s-\phi_2)+\alpha_1\partial_sr_2(s-\phi_1)+\beta_1\beta_2h_0(s-\phi)\\&\quad+\beta_1r_2(s-\phi_1)+\beta_2r_1(s-\phi_2)+\int_{-\rho}^{\rho+2T}r_1(t)\partial_sr_2(s-t)\,dt.\end{aligned}\tag{4.17}$$

When $|s|<2T$ and $x\in B_\rho$, we have $|s-\phi_2|\le 3T$, $|s-\phi_1|\le 3T$. Next, in the integral term in (4.17), we have $|T|<3T$, $-\rho\le s-t\le\rho+2T<3T$. Therefore, in (4.17), the argument of $r_j(t)$, $j=1,2$, always belongs to the interval $|t|\le 3T$; thus we can apply Proposition 3.3 to get

$$\int_{B_\rho}\int_{S^2}\int_{-2T}^{2T}|K(s,\theta,x)|^2ds\,d\theta\,dx\le(C\varepsilon)^2.$$

Therefore, by (4.16), we have

$$\|I_2+I_3+I_4\|_{L^2([-2T,2T]\times S^2)}\le C\varepsilon\|q\|.\tag{4.18}$$

Combining (4.3), (4.7), (4.15), and (4.18), we get

$$\|Pq\|\le C\varepsilon\|q\|.\tag{4.19}$$

On the other hand, by Proposition 4.1, we conclude that

$$C_1\|q\| \le \|Pq\|. \tag{4.20}$$

For ε small enough (4.19) and (4.20) imply $q = 0$. The proof of Theorem 1.1 is complete.

Appendix A. An L^2 estimate. We prove here a theorem for the boundedness of $a(x,y,D)$ in $L^2(\mathbf{R}^n)$ if a is smooth of finite order. Under the assumption that $a = a(x,\xi)$ is independent of y, Theorem 18.1.11′ in [H] says that if $\int|\partial_x^\alpha a(x,\xi)|dx \le M$ for all $\xi\in\mathbf{R}^n$ and for $|\alpha|\le n+1$, then $\|a(x,D)\|_{\mathcal{L}(L^2)} \le CM$ with $C>0$ an absolute constant. Following the proof of that theorem in [H], we obtain a generalization for amplitudes a depending on y as well.

Theorem A.1. *Let A be the operator*

$$Af = (2\pi)^{-n}\iint e^{i(x-y)\cdot\xi}a(x,y,\xi)f(y)\,dy\,d\xi.$$

If

$$\int\left|\partial_x^\alpha\partial_y^\beta a(x,y,\xi)\right|\,dx\,dy \le M \qquad \textit{for } |\alpha|+|\beta|\le 2n+1,\ \xi\in\mathbf{R}^n,$$

then $\|A\|_{\mathcal{L}(L^2)} \le CM$ with $C>0$ an absolute constant.

Proof. We have

$$Af = (2\pi)^{-2n}\iint e^{ix\cdot\xi}\tilde a(x,\xi-\zeta,\xi)\hat f(\zeta)\,d\zeta\,d\xi,$$

where $\tilde a(x,\zeta,\xi) = \int e^{-i\zeta\cdot y}a(x,y,\xi)dy$. Thus

$$\begin{aligned}\widehat{Af}(\eta) := \int e^{-i\eta\cdot x}(Af)(x)\,dx &= (2\pi)^{-2n}\iiint e^{-ix\cdot(\eta-\xi)}\tilde a(x,\xi-\zeta,\xi)\hat f(\zeta)\,d\zeta\,d\xi\,dx\\ &= (2\pi)^{-2n}\iint \tilde{\tilde a}(\eta-\xi,\xi-\zeta,\xi)\hat f(\zeta)\,d\zeta\,d\xi,\end{aligned}$$

where $\tilde{\tilde a}(\eta,\zeta,\xi) = \int e^{-i\eta\cdot x}\tilde a(x,\zeta,\xi) = \int e^{-i(\eta\cdot x+\zeta\cdot y)}a(x,y,\xi)dxdy$. Therefore, $\widehat{Af} = B\hat f$, where B is an integral operator with kernel

$$b(\eta,\zeta) = (2\pi)^{-2n}\int\tilde{\tilde a}(\eta-\xi,\xi-\zeta,\xi)\,d\xi.$$

We claim that $\int|b(\eta,\zeta)|d\eta \le CM$, $\int|b(\eta,\zeta)|d\zeta \le CM$. It is well known that this implies that B is bounded with norm not exceeding CM.

$$\int|b(\eta,\zeta)|\,d\eta \le (2\pi)^{-2n}\iint|\tilde{\tilde a}(\eta-\xi,\xi-\zeta,\xi)|\,d\xi\,d\eta.$$

The assumptions of the theorem imply $|\tilde{\tilde a}(\eta,\zeta,\xi)| \le CM(1+|\eta|+|\zeta|)^{-2n-1}$. Hence

$$\begin{aligned}\int|b(\eta,\zeta)|\,d\eta &\le C'M\iint(1+|\eta-\xi|+|\xi-\zeta|)^{-2n-1}d\eta\,d\xi\\ &= C'M\iint(1+|\eta|+|\xi-\zeta|)^{-2n-1}d\eta\,d\xi\\ &= C'M\iint(1+|\eta|+|\xi|)^{-2n-1}d\eta\,d\xi\\ &= C''M<\infty.\end{aligned}$$

We treat $\int |b(\eta,\zeta)|d\zeta$ in the same way. □

REFERENCES

[B-L-M] A. Bayaliss, Y. Li and C. S. Morawetz,*Scattering by a potential using hyperbolic methods*, Math. Comp., 52 (1989), pp. 321–338.

[C-S] J. Cooper and W. Strauss, *Scattering of waves by periodically moving bodies*, J. Functional. Anal., 47 (1982), pp. 180–229.

[E-R] G. Eskin and J. Ralston, *The inverse backscattering problem in* 3 *dimensions*, Comm. Math. Phys., 124 (1989), pp. 169–215.

[G-U] A. Greenleaf and G. Uhlmann, *Recovering singularities of a potential from singularities of scattering data*, Comm. Math. Phys., 157 (1993), pp. 549–572.

[G] V. Guillemin, *Sojourn times and asymptotic properties of scattering matrix*, Publ. Res. Inst. Math. Sci., 12 (1976), pp. 69–88.

[H] L. Hörmander, *The Analysis of Linear Partial Differential Operators* III, Springer-Verlag, Berlin, 1985.

[L-P] P. Lax and R. Phillips, *Scattering Theory*, Rev. ed., Academic Press, New York, 1989.

[M] C. S. Morawetz, *A formulation for higher dimensional inverse problems for the wave equation*, Comp. Math. Appl., 7 (1981), pp. 319–331.

[N] R. G. Novikov, *Multidimensional inverse spectral problems for the equation* $-\Delta\psi + (v(x) - Eu(x))\psi = 0$, Funct. Anal. Appl., 22 (1988), pp. 263–272.

[P] V. Petkov, *Scattering Theory for Hyperbolic Operators*, North-Holland, Amsterdam, 1989.

[R-S] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 3, Academic Press, New York, 1979.

[St I] P. Stefanov, *A uniqueness result for the inverse back-scattering problem*, Inverse Prob., 6 (1990), pp. 1055–1064.

[St II] P. Stefanov, *Generic uniqueness for two inverse problems in potential scattering*, Comm. Partial Differential Equations, 17 (1992), pp. 55–68.

[S-U] J. Sylvester and G. Uhlmann, *Global uniqueness for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.

[U] G. Uhlmann, *Inverse boundary value problems and applications*, Astérisque, 207 (1992), pp. 153–221.

[V] B. R. Vainberg, *Asymptotic Methods in Equations of Mathematical Physics*, Gordon and Breach, New York, 1988.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1205–1211, September 1997

012

A TWO-POINT PROBLEM WITH NONLINEARITY DEPENDING ONLY ON THE DERIVATIVE*

P. HABETS[†] AND L. SANCHEZ[‡]

Abstract. We study the existence and multiplicity of solutions for a two-point boundary value problem at resonance in the first eigenvalue; the nonlinearity depends only on the first derivative and has finite limits at $\pm\infty$.

Key words. resonance, bounded nonlinearity, multiple solutions

AMS subject classification. 34B15

PII. S003614109529570X

1. Introduction and main result.

The two-point semilinear problem

$$\begin{gathered} u'' + u + g(u') = p(t), \\ u(0) = u(\pi) = 0, \end{gathered} \tag{1}$$

where g is a continuous function in $\mathbb{R}$ and p a continuous function in $[0,\pi]$, has been studied by Cañada and Drábek [3] and Kannan, Nagle, and Pothoven [4]. Related problems involving other boundary conditions are included in [3] and have also been the object of a paper by Mawhin [5].

It has been remarked (see [3], [2], and [4]) that conditions of the Landesman–Lazer type are not appropriated to yield the existence of solutions to (1). Let us denote by $\tilde{C}[0,\pi]$ the subspace of $C[0,\pi]$ consisting of functions $\tilde{u}(t)$ such that $\int_0^\pi \tilde{u}(t)\sin t\,dt = 0$. With respect to the direct sum $C[0,\pi] = \operatorname{span}\{\sin t\} \oplus \tilde{C}[0,\pi]$, every function $p \in C[0,\pi]$ has a decomposition

$$p(t) = \bar{p}\sin t + \tilde{p}(t), \quad \bar{p} \in \mathbb{R},\ \tilde{p} \in \tilde{C}[0,\pi]. \tag{2}$$

In [3] the authors have used a result of Amann, Ambrosetti, and Mancini [1] to show that, given a bounded g, the solvability of (1) can be described in terms of the decomposition of p as follows: for each $\tilde{p}$ there exists a nonempty, bounded interval $I = I(\tilde{p})$ such that (1) has a solution if and only if (2) holds with $\bar{p} \in I$. They compute the upper bound of I in a case where

$$g(-\infty) := \lim_{u\to-\infty} g(u) = g(+\infty) := \lim_{u\to+\infty} g(u).$$

On the other hand, in [4], a contraction argument is used to show that, if $g(s) = \arctan s$, (1) is solvable for p sufficiently small.

In this note, we consider a nonlinearity based on the model studied in [4], and we give a new multiplicity result which in some sense completes the picture given in [3] and [4]; in particular, we prove that there are indeed two solutions of (1) when $\bar{p} \in \overset{\circ}{I}$ and $\bar{p} \neq \frac{2}{\pi}(g(-\infty) + g(+\infty))$. We now state precisely our main result.

[*] Received by the editors December 1, 1995; accepted for publication April 30, 1996. This research was supported by EC contract CHRXCT 940555.

http://www.siam.org/journals/sima/28-5/29570.html

[†] Institut de Mathématique Pure et Appliquée, Chemin du Cyclotron 2, 1348 Louvain la Neuve, Belgium (habets@anma.ucl.ac.be).

[‡] Universidade de Lisboa, Centro de Matemática e Aplicações Fundamentais, Avenida Professor Gama Pinto 2, 1699 Lisboa codex (Lisbon), Portugal (sanchez@ptmat.lmc.fc.ul.pt).

THEOREM 1. *Let $g : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz continuous and such that the limits $g(-\infty)$ and $g(+\infty)$ exist and are finite. Let $p \in C[0,\pi]$ split according to* (2)*. Then there exist real numbers $a = a(\tilde{p}, g) \le b = b(\tilde{p}, g)$ such that*

$$l := \frac{2}{\pi}(g(-\infty) + g(+\infty)) \in [a, b]; \tag{3}$$

and problem (1) *has*

(i) *no solution if $\bar{p} \notin [a,b]$;*
(ii) *at least one solution if $\bar{p} \in (a,b)$ or $\bar{p} \in \{a,b\} \setminus \{l\}$;*
(iii) *at least two solutions if $\bar{p} \in (a,b) \setminus \{l\}$.*

2. Proof of the theorem.

We shall use the following lemma.

LEMMA 2. *Let $m \in L^1(0,a)$ and $v \in W^{2,1}(0,a)$ satisfy the conditions*
(a) *$v'' + m(t)v' + v < 0$ in $(0,a)$,*
(b) *$v(0) = v'(0) = 0$.*
Then there exists $a_1 \in (0,a)$ such that $v(a_1) < 0$.

Proof. Suppose that $v \ge 0$ in $(0,a)$. Let $M(t) = \int_0^t m(s)\,ds$. Then assumption (a) can be written

$$((\exp M(t))v')' + (\exp M(t))v < 0 \quad \text{in } (0,a),$$

and in particular $(\exp M(t))v'(t)$ is strictly decreasing in $[0,a]$. From (b) we infer that $v < 0$ in $(0,a)$, a contradiction. □

Remark. The lemma shows that in fact v takes negative values arbitrarily near $t = 0$.

Proof of the theorem. First we note that without loss of generality we can suppose $l = 0$; i.e., $g(-\infty) + g(+\infty) = 0$. (We shall henceforth add this to our hypotheses.) In fact, letting

$$h(s) := g(s) - \frac{g(-\infty) + g(+\infty)}{2} = g(s) - \frac{\pi l}{4} \quad \text{and} \quad q(t) = p(t) - \frac{\pi l}{4},$$

we obtain an equivalent problem,

$$\begin{gathered} u'' + u + h(u') = q(t), \\ u(0) = u(\pi) = 0. \end{gathered}$$

We also have $a(\tilde{p}, g) - l = a(\tilde{q}, h)$ and $b(\tilde{p}, g) - l = b(\tilde{q}, h)$.

Proof of (i) *and* (ii). The existence of a and b satisfying (i) and (ii) is proved in [3], except possibly for the fact that (1) is solvable when $\bar{p} = b > 0 = l$ or $\bar{p} = a < l = 0$. However, this follows immediately from Claim 1 in the proof of (iii) below, by a standard approximation procedure. □

Proof of (3). Before proceeding to the proof, we shall introduce some notations. Let $C_0^1[0,\pi]$ be the subspace of $C^1[0,\pi]$ consisting of those functions that vanish at $t = 0$ or $t = \pi$ and $\tilde{C}_0^1[0,\pi] = C_0^1[0,\pi] \cap \tilde{C}[0,\pi]$. These are Banach spaces with the usual C^1 norm, $\|u\| := \max(\|u\|_\infty, \|u'\|_\infty)$. Consider the projector $Q : C[0,\pi] \to \tilde{C}[0,\pi]$ defined by

$$Qu(t) = \tilde{u}(t) = u(t) - \frac{2}{\pi}\left(\int_0^\pi u(s) \sin s\,ds\right) \sin t.$$

Denote by K the (linear, compact) inverse of the differential operator $L : \tilde{C}_0^1[0,\pi] \to \tilde{C}[0,\pi]$, $u \mapsto u'' + u$, with domain $\mathrm{Dom}L = C^2[0,\pi] \cap \tilde{C}_0^1[0,\pi]$. The problem

$$\begin{gathered} \tilde{u}'' + \tilde{u} = \tilde{p} - Qg(\bar{u}\cos t + \tilde{u}'), \\ \tilde{u}(0) = \tilde{u}(\pi) = 0 \end{gathered} \tag{4}$$

can be written

$$\tilde{u} = \tilde{T}\tilde{u} := K(\tilde{p} - Qg(\bar{u}\cos t + \tilde{u}')).$$

The operator $\tilde{T} : \tilde{C}_0^1[0,\pi] \to \tilde{C}_0^1[0,\pi]$ is completely continuous and bounded, whence Schauder's theorem applies. This proves that there is a $k > 0$ so that for any given $\bar{u} \in \mathbb{R}$, the problem (4) has a solution $\tilde{u} \in \tilde{C}_0^1[0,\pi]$ with $\|\tilde{u}\| < k$. It follows that (1) has a solution $u(t) = \bar{u}\sin t + \tilde{u}(t)$, for $\bar{p}\sin t = (I - Q)g(\bar{u}\cos t + \tilde{u}')$. By Lebesgue's theorem, we have

$$\lim_{\bar{u}\to\infty} \int_0^\pi g(\bar{u}\cos s + \tilde{u}')\sin s\, ds = 0.$$

Hence, choosing $\bar{u}$ sufficiently large, we obtain numbers

$$\bar{p} := \frac{2}{\pi}\left(\int_0^\pi g(\bar{u}\cos s + \tilde{u}'(s))\sin s\, ds\right)$$

arbitrarily small such that $\bar{p} \in [a,b]$ and, going to the limit, we have $l = 0 \in [a,b]$. □

Proof of (iii). Setting $u(t) = \bar{u}\sin t + \tilde{u}(t)$, with $\bar{u} \in \mathbb{R}$ and $\tilde{u} \in \tilde{C}_0^1[0,\pi]$, it is easily seen that (1) can be rewritten as the system

$$\begin{gathered} \tilde{u} - K[\tilde{p} - Q(g(\bar{u}\cos t + \tilde{u}'))] = 0, \\ \frac{\pi}{2}\bar{p} - \int_0^\pi g(\bar{u}\cos s + \tilde{u}'(s))\sin s\, ds = 0. \end{gathered} \tag{5}$$

As a function of the pair $(\bar{u}, \tilde{u})$, the right-hand side of (5) is a compact perturbation of the identity in $\mathbb{R} \times \tilde{C}_0^1[0,\pi]$; we denote it by $T_{\bar{p}}$ to emphasize its dependence on $\bar{p}$.

We divide this proof into several steps.

Claim 1. Given $\varepsilon > 0$, there exists $R_0 > 0$ such that if u is a solution of (1) for some $\bar{p}$ with $|\bar{p}| \geq \varepsilon$, then $\|u\| < R_0$.

Proof. Assume that there exists a sequence $(\bar{p}_n)_n$ with $|\bar{p}_n| \geq \varepsilon$ and corresponding solutions u_n with $\|u_n\| \to \infty$. Split $u_n(t) = \bar{u}_n \sin t + \tilde{u}_n(t)$ according to (2). As in the above argument, it follows from (5) that the sequence $\|\tilde{u}_n\|$ is bounded. Hence $|\bar{u}_n| \to \infty$. Also, we obtain from (5)

$$\left|\int_0^\pi g(\bar{u}_n\cos s + \tilde{u}_n'(s))\sin s\, ds\right| = \frac{\pi}{2}|\bar{p}_n| \geq \frac{\pi}{2}\varepsilon,$$

and, by Lebesgue's theorem again, we reach the contradiction $0 \geq \frac{\pi}{2}\varepsilon$. □

Claim 2. Set $B_R = \{(\bar{u}, \tilde{u}) \mid u = \bar{u}\sin t + \tilde{u} \in C_0^1[0,\pi], \|u\| \leq R\}$. Given $\bar{p}_0 \neq 0$, there exists R_0 such that for every $R > R_0$,

$$\deg(T_{\bar{p}_0}, B_R, 0) = 0. \tag{6}$$

Proof. Suppose $\bar{p}_0 > 0$. As shown in Claim 1, there exists R_0 such that for all solutions u of (5) with $\bar{p} \geq \bar{p}_0$, we have $\|u\| < R_0$. On the other hand, by (i), we can fix $\bar p_1 > \bar p_0$ such that (5) (with $\bar p = \bar p_1$) has no solution at all. The homotopy and existence properties of Leray–Schauder degree allow us to conclude that for every $R > R_0$,

$$\deg(T_{\bar p_0}, B_R, 0) = \deg(T_{\bar p_1}, B_R, 0) = 0.$$

The same argument applies if $\bar p_0 < 0$. $\square$

Claim 3. If $0 < \bar p < b$ or $0 > \bar p > a$, there exists a bounded open set $\Omega \subset \mathbb{R} \times \tilde C_0^1[0,\pi]$ such that

$$|\deg(T_{\bar p}, \Omega, 0)| = 1.$$

Proof. Step 1*: Construction of the set* Ω. To fix ideas, suppose that $0 < \bar p < b$. Choose $\rho \in (\bar p, b)$. By the characterization of a and b, there exists a solution of the problem

$$\begin{gathered}\alpha'' + \alpha + g(\alpha') = \rho \sin t + \tilde p(t) > \bar p \sin t + \tilde p(t), \qquad \text{on } (0,\pi),\\ \alpha(0) = \alpha(\pi) = 0.\end{gathered} \tag{7}$$

Fix $0 < \varepsilon < \frac{\pi}{4}\bar p$. Recall that there is $k > 0$ so that any solution of (5) is such that $\|\tilde u(t)\| \le k$. Let us then choose $\bar\beta_0$ large enough so that for any $\tilde u$ with $\|\tilde u(t)\| \le k$,

$$\bar\beta_0 \sin t > \alpha(t) + \tilde u(t) \quad \text{ in } (0,\pi)$$

and

$$\begin{aligned}(I-Q)g(\bar\beta_0\cos t + \tilde u') &= \frac{2}{\pi}\left(\int_0^\pi g(\bar\beta_0 \cos s + \tilde u'(s)) \sin s\, ds\right)\sin t\\ &< \left(\bar p - \frac{4\varepsilon}{\pi}\right)\sin t.\end{aligned}$$

Next, we choose a $\tilde\beta_0$ solution of

$$\tilde u - K[\tilde p - \tilde\varepsilon - Q(g(\bar\beta_0 \cos t + \tilde u'))] = 0,$$

where $\tilde\varepsilon = Q\varepsilon = \varepsilon(1 - \frac{4}{\pi}\sin t)$. The function $\beta_0(t) := \bar\beta_0 \sin t + \tilde\beta_0(t)$ satisfies

$$\begin{aligned}\beta_0'' + \beta_0 + g(\beta_0') &= \tilde p - \tilde\varepsilon + (I-Q)g(\bar\beta_0\cos t + \tilde\beta_0')\\ &< \tilde p + \bar p \sin t - \varepsilon, \qquad\qquad 0 < t < \pi,\\ &\beta_0(0) = \beta_0(\pi) = 0,\end{aligned}$$

and therefore $\beta(t) := \beta_0(t) + \varepsilon$ is a strict upper solution of (1) such that $\beta \ge \alpha + \varepsilon$ in $[0,\pi]$.

Choose $k > 0$ so large that

$$k(\beta - \alpha) > \beta'' - \alpha'' \quad \text{in } [0,\pi]. \tag{8}$$

Then let us consider the homotopy

$$\begin{gathered}u'' + \lambda(u + g(u') - p) - (1-\lambda)\left[k\left(u - \tfrac{\alpha+\beta}{2}\right) + \tfrac{\alpha''+\beta''}{2}\right] = 0,\\ u(0) = u(\pi) = 0, \qquad\qquad 0 \le \lambda \le 1,\end{gathered} \tag{9}$$

and the bounded open subset of $\mathbb{R} \times \tilde{C}_0^1[0,\pi]$,

$$\begin{aligned}\Omega = \{(\bar{u}, \tilde{u}) \mid u = \bar{u}\sin t + \tilde{u} \in C_0^1[0,\pi], \alpha(t) < u(t) < \beta(t) \text{ in } (0,\pi), \\ \|u'\|_\infty < N,\ u'(0) > \alpha'(0) \ \text{ and } \ u'(\pi) < \alpha'(\pi)\},\end{aligned}$$

where N is a bound for derivatives of solutions u of (9) such that $\alpha \le u \le \beta$. This bound clearly exists, since g is bounded.

Step 2*: There exists no solution $u = \bar{u}\sin t + \tilde{u}$ of* (9) *such that $(\bar{u}, \tilde{u}) \in \partial\Omega$.* First, we prove that there exists no solution $u = \bar{u}\sin t + \tilde{u}$ of (9) on $\partial\Omega$ if $0 \le \lambda < 1$. Arguing by contradiction, suppose that for some $s \in [0,\pi]$ we have $u(s) = \alpha(s)$, $u'(s) = \alpha'(s)$, $u''(s) \ge \alpha''(s)$. Then we compute

$$\alpha''(s) \le u''(s) \le \lambda\alpha''(s) + (1-\lambda)\left[k\frac{\alpha(s)-\beta(s)}{2} + \frac{\alpha''(s)+\beta''(s)}{2}\right]$$

so that

$$\alpha''(s) \le k\frac{\alpha(s)-\beta(s)}{2} + \frac{\alpha''(s)+\beta''(s)}{2},$$

contradicting the choice of k in (8). An analogous computation shows that it is impossible to have $u(s) = \beta(s)$, $u'(s) = \beta'(s)$, $u''(s) \le \beta''(s)$ for some $s \in (0,\pi)$.

Next, we show that, if $\lambda = 1$, no solution $u \in \overline{\Omega}$ can satisfy $u(s) = \alpha(s)$, $u'(s) = \alpha'(s)$ or $u(s) = \beta(s)$, $u'(s) = \beta'(s)$. This is straightforward since, in case $0 < s < \pi$,

$$\alpha''(s) + \alpha(s) + g(\alpha'(s)) - p(s) > 0 > \beta''(s) + \beta(s) + g(\beta'(s)) - p(s).$$

Also, $\beta(0) = \beta(\pi) = \varepsilon > 0$. Hence, it remains to show that $u(s) = \alpha(s)$, $u'(s) = \alpha'(s)$ cannot hold with $(\bar{u}, \tilde{u}) \in \overline{\Omega}$ and $s = 0$ or $s = \pi$. Let us consider the case $s = 0$, the other being similar. If $u(0) = \alpha(0)$, $u'(0) = \alpha'(0)$, the function $v(t) := u(t) - \alpha(t)$ would satisfy $v \ge 0$ in $[0,\pi]$ and also, on account of (7),

$$v'' + v + m(t)v' = (\bar{p} - \rho)\sin t < 0 \quad \text{in } [0,\pi],$$

where $m(t) := \frac{g(u'(t)) - g(\alpha'(t))}{u'(t) - \alpha'(t)}$ if $u'(t) \ne \alpha'(t)$ and $m(t) = 0$ otherwise. Since g is locally Lipschitz continuous, the function m is (measurable and) bounded, and we obtain a contradiction with Lemma 2.

Step 3*: Proof of the claim.* Note that (9) can be written in operator form as

$$\begin{aligned}\tilde{u} - KQ[\lambda N(\bar{u},\tilde{u}) + (1-\lambda)M(\bar{u},\tilde{u})] = 0, \\ \int_0^\pi [\lambda N(\bar{u},\tilde{u})(t) + (1-\lambda)M(\bar{u},\tilde{u})(t)]\sin t\, dt = 0,\end{aligned}$$

where

$$\begin{aligned}N(\bar{u},\tilde{u})(t) = p - g(\bar{u}\cos t + \tilde{u}'(t)) \quad \text{and} \\ M(\bar{u},\tilde{u})(t) = k\left(\bar{u}\sin t + \tilde{u}(t) - \tfrac{\alpha(t)+\beta(t)}{2}\right) + \tfrac{\alpha''(t)+\beta''(t)}{2} - \bar{u}\sin t - \tilde{u}(t).\end{aligned}$$

For $\lambda = 1$, this reduces to $T_{\bar{p}} = 0$. On the other hand, for $\lambda = 0$, it is clear that this is an equation of the form $L = 0$, where L is a linear, invertible compact perturbation of the identity. We derive from (9) that its unique solution is

$$u(t) = \frac{\alpha(t)+\beta(t)}{2} - \frac{\varepsilon}{2}\frac{\cosh\sqrt{k}(t-\pi/2)}{\cosh\sqrt{k}\pi/2} = \bar{u}\sin t + \tilde{u}(t),$$

and it is clear that $(\bar{u}, \tilde{u})$ belongs to Ω. By the invariance property of Leray–Schauder degree, we easily conclude the proof of Claim 3. $\square$

Claim 4. Assertion (iii) holds.

With $0 < \bar{p} < b$ fixed, we construct Ω as in Claim 3 and $R > R_0$ given by Claim 2 so that $\Omega \subset B_R$. Then the excision property of Leray–Schauder degree implies

$$|\deg(T_{\bar{p}}, B_R \setminus \Omega, 0)| = 1.$$

The existence of a solution of (5) (with $\bar{p} = \bar{p}_0$) in $B_R \setminus \Omega$ follows. There exists another solution in Ω by Claim 3. Similarly, we conclude that there exist two solutions if $a < \bar{p} < 0$, and the proof is complete. $\square$

3. Additional results. We do not know whether the interval $[a(\tilde{p}, g), b(\tilde{p}, g)]$ is not reduced to a point for some $\tilde{p}$. However, following the idea of [4], we can give simple conditions to ensure that, at least for small $\tilde{p}$, it is in fact nondegenerate.

PROPOSITION 3. *Assume that* $g : \mathbb{R} \to \mathbb{R}$ *is of class* C^1, $g(0) = 0$, *the limits* $g(-\infty)$ *and* $g(+\infty)$ *exist, and* $g'(0) \neq 0$. *Let* $p \in C[0, \pi]$ *split according to* (2). *Then there exists* $\varepsilon > 0$ *such that, if* $\|p\|_\infty < \varepsilon$, *we have* $a(\tilde{p}, g) < 0 < b(\tilde{p}, g)$.

Proof. The mapping $\mathcal{F} : C^2[0, \pi] \cap C_0^1[0, \pi] \to C[0, \pi]$, $u \mapsto u'' + u + g(u')$ is differentiable at $u = 0$. Since the problem

$$\begin{gathered} v'' + v + cv' = 0, \\ v(0) = v(\pi) = 0, \end{gathered}$$

where c is a nonzero constant, has only the trivial solution, it follows that $\mathcal{F}'(0)$ is an isomorphism. Hence, by the inverse mapping theorem, the range of $\mathcal{F}$ contains an open ball centered at the origin, and we can conclude. $\square$

Another interesting feature of problem (1) under the assumptions of the above proposition is that for p small and $\bar{p} \neq 0 = l$, one can indeed prove the existence of two *ordered* solutions.

PROPOSITION 4. *Assume that* $g : \mathbb{R} \to \mathbb{R}$ *is of class* C^1, $g(0) = 0$, *the limits* $g(-\infty)$ *and* $g(+\infty)$ *exist, and* $g'(0) \neq 0$. *Let* $p \in C[0, \pi]$ *split according to* (2) *and* $\bar{p} \neq l = 0$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|p\|_\infty < \varepsilon$, *we have two solutions* u_1 *and* u_2 *of* (1) *with*

$$u_1(t) > u_2(t), \qquad \textit{in } (0, \pi).$$

Proof. Let $\|p\|_\infty$ be so small that the corresponding solution $u_1(t)$ given by the inverse mapping theorem has the property that the solution of the linear Cauchy problem

$$v'' + v + g(u_1')v' = 0, \quad v(0) = 0,\ v'(0) = 1 \tag{10}$$

is *positive* in $(0, \pi]$. Making the change of variable $u = u_1(t) + w$ in (1), we obtain a new problem

$$w'' + w + g(u_1'(t) + w') = g(u_1'(t)), \quad w(0) = w(\pi) = 0. \tag{11}$$

For definiteness, assume $\bar{p} > 0$. Consider the solution $z(t, \lambda)$ of the Cauchy problem associated with (11):

$$z'' + z + g(u_1'(t) + z') = g(u_1'(t)), \quad z(0) = 0,\ z'(0) = \lambda.$$

Since $v(t) := \frac{\partial z}{\partial \lambda}(t,0)$ is the solution of (10), we deduce that for $\lambda < 0$ and $|\lambda|$ sufficiently small, we have $z(t,\lambda) < 0$ for every $t \in (0,\pi]$. On the other hand, substituting $z(t,\lambda) = \lambda \sin t + x(t,\lambda)$, $x(t,\lambda)$ solves the problem

$$x'' + x + g(u_1'(t) + \lambda \cos t + x') = g(u_1'(t)), \quad x(0) = x'(0) = 0.$$

As $\lambda \to -\infty$, $x(t,\lambda)$ converges uniformly in $[0,\pi]$ to the solution $\hat{x}(t)$ of

$$x'' + x + \hat{g}(t) = g(u_1'(t)), \quad x(0) = x'(0) = 0,$$

where $\hat{g}(t) = g(-\infty)$ if $0 < t < \frac{\pi}{2}$ and $\hat{g}(t) = g(+\infty)$ if $\frac{\pi}{2} < t < \pi$. Now this is given explicitly by

$$\hat{x}(t) = \int_0^t \sin(t-s)[g(u_1'(s)) - \hat{g}(s)]\, ds.$$

As $\hat{x}(\pi) = 2\bar{p} > 0$, we have $x(\pi,\lambda) > 0$ for $|\lambda|$ large. We infer that for some (negative) value of λ, $z(.,\lambda)$ is a negative solution of (11). Hence, (1) has a second solution $u_2(t)$ such that $u_2(t) < u_1(t)$ in $(0,\pi)$. □

REFERENCES

[1] H. Amann, A. Ambrosetti, and G. Mancini, *Elliptic equations with noninvertible Fredholm linear part and bounded nonlinearities*, Math. Z., 158 (1978), pp. 179–194.

[2] A. Cañada, *Existencia de soluciones para problemas de contorno elipticos no lineales y no necesariamente autoadjuntos*, VII C.E.D.Y.A., Granada, 1984, pp. 79–84.

[3] A. Cañada and P. Drábek, *On semilinear problems with nonlinearities depending only on derivatives*, SIAM J. Math. Anal., 27 (1996), pp. 543–557.

[4] R. Kannan, R. Kent Nagle, and K. L. Pothoven, *Remarks on the existence of solutions of* $x'' + x + \arctan(x') = p(t)$*,* $x(0) = x(\pi) = 0$, Nonlinear Anal., 22 (1994), pp. 793–796.

[5] J. Mawhin, *Some remarks on semilinear problems at resonance where the nonlinearity depends only on the derivatives*, Acta Math. Inform. Univ. Ostraviensis, 2 (1994), pp. 61–69.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1212–1226, September 1997

MEASURABLE MULTIFUNCTIONS IN NONSEPARABLE BANACH SPACES [*]

DIÓMEDES BÁRCENAS[†] AND WILFREDO URBINA[‡]

Abstract. In this article we define measurable multifunctions in nonseparable Banach spaces, prove a weak compactness criterion for the selectors of multifunctions integrably bounded, characterize Banach spaces that have the Radon–Nikodym property by means of convergence of multivalued martingales, generalize some recent results on convergence of set-valued conditional expectations, and give some applications to control theory and differential inclusions.

Key words. measurable multifunctions, set-valued martingales, multiconditional expectation, Radon–Nikodym property, Kuratowski–Mosco convergence, controllability, differential inclusions

AMS subject classifications. 28B05, 47D06, 93B05

PII. S0036141095296005

1. Introduction.

Beyond being interesting by itself, the theory of measurable multifunctions has been shown to be useful in many branches of mathematics, such as convex analysis [4], mathematical economy [20], differential inclusions [13], and control theory [1].

In this paper we start with the definition of measurable multifunctions in nonseparable Banach spaces, then give a compactness criterion for the set of all measurable selections of an integrably bounded multifunction and a characterization of Banach spaces with the Radon–Nikodym property. Also, we generalize some results related to the convergence of multiconditional expectations recently obtained by Papageorgiou [23], prove a sort of continuous dependence theorem of the attainable set for a suitable infinite-dimensional linear system in terms of Kuratowski–Mosco convergence, and, finally, provide conditions for the compactness of mild trajectories of linear differential inclusions.

We want to remark that most of these results work in general Banach spaces, dropping some hypotheses like separability and reflexivity.

2. Preliminaries.

Let (Ω, Σ, μ) be a nonnegative, complete, σ-finite measure space. For a Banach space X we shall use the following notation:

$$\mathbf{P}_{f(c)}(X) = \{A \subset X : A \neq \emptyset, \text{closed (convex)}\},$$

$$\mathbf{P}_{wkc}(X) = \{A \subset X : A \neq \emptyset, \text{weakly compact, convex}\}.$$

For a nonempty subset A of X we put $|A| = \sup_{x \in A} ||x||$.

It has been standard to define *measurable multifunctions* with values in a separable Banach space as follows. Given a separable Banach space X and a measurable space

[*] Received by the editors December 13, 1995; accepted for publication April 30, 1996.
http://www.siam.org/journals/sima/28-5/29600.html

[†] Departamento de Matemática, Facultad de Ciencias, Universidad de los Andes, Mérida 5101, Venezuela (barcenas@ciens.ula.ve). The research of this author was supported by CDCHT-ULA project C-593-93.

[‡] Departamento de Matemática, Facultad de Ciencias, Universidad Central de Venezuela, Apartado 47195 Los Chaguaramos, Caracas 1041-A, Venezuela (wurbina@klein.ciens.ucv.ve).

(Ω, Σ), a multifunction $\mathbf{F} : \Omega \to 2^X \backslash \{\emptyset\}$ is measurable if for each open subset U of X we have

$$\mathbf{F}^-(U) = \{\omega \in \Omega : \mathbf{F}(\omega) \cap U \neq \emptyset\} \in \Sigma.$$

In [4] the following result is proved.

THEOREM 2.1. *Let* $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ *be a measurable multifunction. If there is a nonnegative complete, σ-finite measure μ defined on Σ, the following statements are equivalent:*

(i) $\mathbf{F}$ *is measurable.*

(ii) *For every* $x \in X,\ \omega \to d(x, \mathbf{F}(\omega))$ *is measurable.*

(iii) *There is a sequence of measurable functions* $f_n : \Omega \to X$ *such that* $\mathbf{F}(\omega) = \overline{\{f_n(\omega)\}}, \forall \omega \in \Omega$ *(Castaing representation).*

(iv) $\mathbf{GrF} = \{(\omega, x) : x \in \mathbf{F}(\omega)\} \in \Sigma \times \mathbf{B}(X)$*, with* $\mathbf{B}(X)$ *the Borel σ-field of X.*

(v) $\mathbf{F}^-(C) = \{\omega \in \Omega : \mathbf{F}(\omega) \cap C \neq \emptyset\} \in \Sigma$ *for each C closed subset of X.*

We notice that the *separability condition* in the Banach space X is a very strong restriction which can be removed thanks to the Castaing representation together with the Pettis measurability criterion ([8, p. 42]). This will be done in the next section.

Given a measurable multifunction $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$, a selector of $\mathbf{F}$ is a measurable function $f : \Omega \to X$ such that $f(\omega) \in \mathbf{F}(\omega)$, μ a.e.

We will denote $\mathbf{S}^1_{\mathbf{F}}$ as the collection of selectors of $\mathbf{F}$ that are Bochner integrable. Given that definition, the following result holds [19]. $\mathbf{S}^1_{\mathbf{F}} \neq \emptyset$ if and only if the function $g : \Omega \to X$ defined by

$$g(\omega) = \inf\{||x||; x \in \mathbf{F}(\omega)\}$$

is Lebesgue integrable.

By using $\mathbf{S}^1_{\mathbf{F}}$, we can define the multivalued integral

$$\int_\Omega \mathbf{F} d\mu = \left\{ \int_\Omega f d\mu : f \in \mathbf{S}^1_{\mathbf{F}} \right\}.$$

A multifunction $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ is called *integrably bounded* if there is an $f \in L^1(\mu)$ such that $|\mathbf{F}(\omega)| \leq f(\omega)$, μ. a.e.

A collection $\{\mathbf{F}_\alpha\}_{\alpha \in \Lambda}$ of measurable multifunctions is called *uniformly integrable* if

(i) $\{|\mathbf{F}_\alpha|\}_{\alpha \in \Lambda}$ is bounded in $L^1(\mu)$, and

(ii) given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\mu(E) < \delta \Rightarrow \int_E |\mathbf{F}_\alpha| d\mu < \varepsilon$$

for each $\alpha \in \Lambda$.

Given a complete probability space $(\Omega, \Sigma, \mathbf{P})$, an integrably bounded multifunction $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$, and a sub-$\sigma$-field $\Sigma_0 \subset \Sigma$, we define the *multiconditional expectation of* $\mathbf{F}$ *with respect to* Σ_0 as the multifunction $\mathbf{E}^{\Sigma_0}_{\mathbf{F}} : \Omega \to \mathbf{P}_f(X)$, for which we have

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{F}}} = \overline{\{\mathbf{E}^{\Sigma_0}_{\mathbf{f}} : f \in \mathbf{S}^1_{\mathbf{F}}\}},$$

where the closure is taken in the norm topology of $L^1_X(\mu)$. It has been proved by Hiai and Umegaki ([15, Cor. 1.2]) that $\mathbf{E}^{\Sigma_0}_{\mathbf{F}}$ exists and is unique.

If $(\Omega, \Sigma, \mathbf{P})$ is a complete probability space, $\{\Sigma_n\}_{n\in\mathbf{N}}$ is an increasing sequence of sub-σ-fields of Σ, and $\mathbf{F}_n : \Omega \to \mathbf{P}_f(X)$ is a sequence of integrably bounded multifunctions such that for each $n \in \mathbf{N}$, $\mathbf{F}_n$ is $\mathbf{P}$-measurable with respect to Σ_n, we say that $(\mathbf{F}_n, \Sigma_n)_{n\geq 1}$ is a *set-valued martingale* if

$$\mathbf{E}^{\Sigma_n}_{\mathbf{F}_{n+1}}(\omega) = \mathbf{F}_n(\omega) \text{ a.s.}$$

In what follows we shall also need the following definition. We say that a sequence $\{\Sigma_n\}_{n\in\mathbf{N}}$ of sub-σ-fields of Σ *converges to* Σ_0 in $L^1_X(\mathbf{P})$ if for each $f \in L^1_X(\mathbf{P})$, $\mathbf{E}^{\Sigma_n}_f \to \mathbf{E}^{\Sigma_0}_f$ in the norm topology of $L^1_X(\mathbf{P})$.

Finally, following Mosco [17], for a sequence $\{A_n\}_{n\geq 1}$ contained in $2^X\backslash\{\emptyset\}$, we set

$$\begin{aligned} s - \underline{\lim} A_n &= \{x \in X : x = \lim_{n\to\infty} x_n, x_n \in A_n\} \\ &= \{x \in X : x = \lim_{n\to\infty} d(x, A_n) = 0\} \end{aligned}$$

and

$$\omega - \overline{\lim} A_n = \{\omega - \lim_{k\to\infty} x_k, x_k \in A_{n_k}$$

for some subsequence $\{A_{n_k}\}$ of $\{A_n\}\}$.

According to Papageorgiou [23], a sequence $\{A_n\}$ is said to be convergent *in the Kuratowski–Mosco sense* to a set A (that we will denote as $A_n \xrightarrow{K.M.} A$) if

$$s - \overline{\lim} A_n = \omega - \overline{\lim} A_n = A.$$

3. Measurable multifunctions. In this section we start removing separability in the definition of measurable multifunctions.

Definition 3.1. *Let X be an arbitrary Banach space and (Ω, Σ, μ) be a positive, σ-finite complete measure space. A multifunction $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ is μ-measurable if there is a sequence of μ-measurable functions $f_n : \Omega \to X$ and $N \in \Sigma$ with $\mu(N) = 0$ so that*

$$\mathbf{F}(\omega) = \overline{\{f_n(\omega)\}}, \quad \forall \omega \in \Omega\backslash N.$$

With this definition, we restate Theorem 2.1 in the following fashion.

Theorem 3.1. *Let X be a Banach space and $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ a multifunction. The following conditions are equivalent:*

(i) $\mathbf{F}$ *is measurable.*

(ii) *There is an $N \in \Sigma$ with $\mu(N) = 0$ such that $Y = [\cup_{\omega\in\Omega\backslash N}\mathbf{F}(\omega)]$, the closed subspace generated by $\cup_{\omega\in\Omega\backslash N}\mathbf{F}(\omega)$ is separable, and for each relatively open set $U \subset Y$, $\overline{\mathbf{F}}(U) \in \Sigma$.*

(iii) *There is an $N \in \Sigma$ with $\mu(N) = 0$ such that $Y = [\cup_{\omega\in\Omega\backslash N}\mathbf{F}(\omega)]$ is separable, and for each $z \in Y$, the function $g : \Omega \to R$ defined by*

$$g(\omega) = \begin{cases} 0, & \text{if } \omega \in N, \\ d(z, \mathbf{F}(\omega)), & \text{if } \omega \in \Omega\backslash N, \end{cases}$$

is μ-measurable.

(iv) *There is an* $N \in \Sigma$ *with* $\mu(N) = 0$ *such that* $Y = [\bigcup_{\omega \in \Omega\backslash N} \mathbf{F}(\omega)]$ *is separable and*

$$\mathbf{GrF} = \{(\omega, x) : x \in \mathbf{F}(\omega)\} \in \Sigma \times \mathbf{B}(Y),$$

where $\mathbf{B}(Y)$ *is the Borel* σ*-field of* Y*.*

(v) *There is an* $N \in \Sigma$ *with* $\mu(N) = 0$ *such that* $Y = \overline{[\bigcup_{\omega \in \Omega\backslash N} \mathbf{F}(\omega)]}$ *is separable and* $\mathbf{F}^{-}(C) \in \Sigma$ *for every* C *relatively closed in* Y*.*

Proof. Since it relies on the separable case, we will prove the equivalence between (i) and (ii) in order to illustrate the technique.

(i)⇒(ii) Since $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ is μ-measurable, there is a sequence of μ-measurable functions $f_n : \Omega \to X$ and there is a set $N_0 \in \Sigma$ so that $\mu(N_0) = 0$ and

$$\mathbf{F}(\omega) = \overline{\{f_n(\omega)\}}, \forall \omega \in \Omega\backslash N_0.$$

Since each f_n is μ-measurable, the Pettis μ-measurability criterion implies that for each $n \in \mathbf{N}$ there is an $N_n \in \Sigma$ such that $\mu(N_n) = 0$ and $\overline{f_n(\Omega\backslash N_n)}$ is separable.

If $N = \bigcup_{n=0}^{\infty} N_n$, then $N \in \Sigma$, $\mu(N) = 0$, and $f_n(\Omega\backslash N)$ is separable, which implies that $Y = [\bigcup_{\omega \in \Omega\backslash N} \mathbf{F}(\omega)]$ is separable.

Define $\mathbf{G} : \Omega \to \mathbf{P}_f(Y)$ by

$$\mathbf{G}(\omega) = \begin{cases} \mathbf{F}(\omega), & \text{if } \omega \in \Omega\backslash N, \\ \{0\}, & \text{if } \omega \in N, \end{cases}$$

and $g_n : \Omega \to X$ by

$$g_n(\omega) = \begin{cases} f_n(\omega), & \text{if } \omega \in \Omega\backslash N, \\ 0, & \text{if } \omega \in N. \end{cases}$$

Since μ is complete, each g_n is measurable and

$$G(\omega) = \overline{\{g_n(\omega)\}} \;\; \forall \omega \in \Omega.$$

Therefore, G is a μ-measurable multifunction. This fact, together with the separability of Y and the completeness of μ, shows that the equivalence (i)–(ii) of Theorem 2.1 holds with $\mathbf{G}$ instead of $\mathbf{F}$, and the conclusion follows.

Now suppose (ii) holds and define $\mathbf{G} : \Omega \to \mathbf{P}_f(Y)$ as before. Since Y is separable, μ is complete, and $\mathbf{G}$ is plainly μ-measurable, we get, by Theorem 2.1, that there is a sequence $\{f_n\}$ of measurable functions $f_n : \Omega \to X$ such that $\mathbf{G}(\omega) = \overline{\{f_n(\omega)\}}$ for each $\omega \in \Omega$. Since $\mathbf{G}(\omega) = \mathbf{F}(\omega), \forall \omega \in \Omega\backslash N$, then $\mathbf{F}(\omega) = \overline{\{f_n(\omega)\}}$ $\forall \omega \in \Omega\backslash N$. Thus $\mathbf{F}$ is μ-measurable. □

Now, in this case, the definitions of $\mathbf{S}^1_{\mathbf{F}}$, multiconditional expectation, set-valued martingale, and all the other notions are analogous to the separable case.

Our next goal is to give a characterization of weak compactness in $L^1_X(\mu)$ of $\mathbf{S}^1_{\mathbf{F}}$ for $\mathbf{F}$ integrably bounded.

Theorem 3.2. *Let* $\mathbf{F} : \Omega \to \mathbf{P}_{fc}(X)$ *be an integrably bounded multifunction with* μ *finite, nonnegative, and complete. Then* $\mathbf{S}^1_{\mathbf{F}}$ *is weakly compact if and only if* $\mathbf{F}(\omega)$ *is weakly compact,* μ *a.e.*

Proof. (⇐) Since the multifunction $\mathbf{F}$ is integrably bounded, $\mathbf{S}^1_{\mathbf{F}}$ is uniformly integrable. Since for almost all $\omega \in \Omega, \mathbf{F}(\omega)$ is weakly compact, the set

$$\{f : f \in \mathbf{S}^1_{\mathbf{F}}\}$$

is relatively weakly compact in X. Now consider a sequence $\{f_n\}$ in $\mathbf{S}^1_{\mathbf{F}}$. By Ülger [27] and Diestel, Ruess, and Schachermayer [7], there is a $g_n : \Omega \to X$ so that $g_n(\omega) \in c_0\{f_n(\omega), f_{n+1}(\omega), \ldots\}$ for μ (a.e.) ω and $g \in L^1_X(\mu)$ so that

$$g_n(\omega) \to g(\omega)$$

(μ a.e.) in the norm of X with $g(\omega) \in \mathbf{F}(\omega), \mu$ a.e. Consequently, $\mathbf{S}^1_{\mathbf{F}}$ is relatively weakly compact. Plainly, $\mathbf{S}^1_{\mathbf{F}}$ is a convex subset of $L^1_X(\mu)$.

Now let $f \in \overline{\mathbf{S}^1_{\mathbf{F}}}$. Then there exists a sequence $\{f_n\}_n$ in $\mathbf{S}^1_{\mathbf{F}}$ so that $\{f_n\}_n$ converges to f in the strong topology of $L^1_X(\mu)$; hence, there is a subsequence $\{f_{n_k}\}_k$ of $\{f_n\}_n$ satisfying $f_{n_k}(\omega) \to f(\omega)$, μ a.e.

Since $\{f_{n_k}\}_k \subset \mathbf{F}(\omega)$, which is convex and weakly compact (μ a.e.), hence closed in the strong topology of X, we get $f(\omega) \in \mathbf{F}(\omega)$, μ a.e. This implies that $\mathbf{S}^1_{\mathbf{F}}$ is closed and then, by convexity, weakly closed.

($\Rightarrow$) Suppose $\mathbf{F}(\omega)$ is not weakly compact (μ a.e.); then there is a measurable set A with $\mu(A) > 0$, and $\mathbf{F}(\omega)$ is not relatively weakly compact for $\omega \in A$. Since $\mathbf{F}$ is integrably bounded, $\mathbf{S}^1_{\mathbf{F}}$ coincides with the set of all measurable selectors of $\mathbf{F}$ and, again applying the Ülger–Diestel–Ruess–Schachermayer theorem, $\mathbf{S}^1_{\mathbf{F}}$ is not relatively weakly compact, which is a contradiction. So, $\mathbf{F}(\omega)$ has to be relatively weakly compact (μ a.e.) and, being closed, weakly compact, μ a.e. □

Our proof not only generalizes the proof given by Papageorgiou in [18, 19] but is more elementary than that one. In fact, Papageorgiou's proof relies on the well-known James compactness criterion [16]; our proof relies instead on the Ülger–Diestel–Ruess–Schachermayer $L^1_X(\mu)$ compactness criterion, which uses Mazur's theorem, which is more elementary than James' theorem. Furthermore, our proof is close in spirit to the theory of single-valued functions.

We end this section with a characterization of the Radon–Nikodym property (compare with Egghe [11, Thm. II.2.2.1]).

Theorem 3.3. *For a Banach space X the following statements are equivalent:*

(i) *X has the Radon–Nikodym property.*

(ii) *For every set-valued uniformly integrable martingale $\mathbf{F}_n : \Omega \to \mathbf{P}_f(X)$, there is an integrably bounded multifunction $\mathbf{F} : \Omega \to \mathbf{P}_{fc}(X)$ such that $\mathbf{E}^{\Sigma_n}_{\mathbf{F}}(\omega) = \mathbf{F}_n(\omega)$, μ a.e.*

(iii) *For every L^∞-bounded set-valued martingale $\mathbf{F}_n : \Omega \to \mathbf{P}_{fc}(X)$, there is an integrably bounded multifunction $\mathbf{F} : \Omega \to \mathbf{P}_{fc}(X)$ such that $\mathbf{E}^{\Sigma_n}_{\mathbf{F}}(\omega) = \mathbf{F}_n(\omega)$, μ a.e.*

Proof. (i) $\Rightarrow$ (ii) Since $\{\mathbf{F}_n\}_n$ is a sequence of measurable multifunctions, we can suppose, without loss of generality, that the closed subspace Y generated by $\{\bigcup_n \bigcup_{\omega\in\Omega} \mathbf{F}_n\}$ is separable. With this assumption, we have that

$$\mathbf{F}_n : \Omega \to \mathbf{P}_{fc}(Y).$$

Since the Radon–Nikodym property is hereditary for closed subspaces, the conclusion follows using Theorem 3.1 of [22].

(ii) $\Rightarrow$ (iii) It is trivial.

(iii) $\Rightarrow$ (i) If (iii) holds, it does for single-valued martingales, and this leads to the well-known classical case (see Van Dust [10]). □

4. Convergence of multiconditional expectation. In this section we generalize Theorems 3.1 and 3.2 of [23] in several directions. The first generalization is the following one.

THEOREM 4.1. *If X^* has the Radon–Nikodym property, $\mathbf{F} : \Omega \to \mathbf{P}_{fc}(X)$ is an integrably bounded multifunction and Σ_n an increasing sequence of sub-σ-fields converging to Σ_0, then*

$$\mathbf{E}_{\mathbf{F}}^{\Sigma_n}(\omega) \xrightarrow{K.M.} \mathbf{E}_{\mathbf{F}}^{\Sigma_0}(\omega),$$

μ a.e. The result is also true if X^* has the Radon–Nikodym property and (Ω, Σ, μ) has no Σ_0-atoms.

Proof. If X^* is separable, the result is Theorem 3.1 of [23]. In the general case, by Stegall's result [25], X^* has the Radon–Nikodym property only if every separable subspace of X has a separable dual. Since $\mathbf{F}$ is μ-measurable, there is an $N \in \Sigma$ such that $\mu(N) = 0$ and Y, the closed subspace generated by $\bigcup_{\omega \in \Omega \setminus N} \mathbf{F}(\omega)$, is separable. Define $\mathbf{G} : \Omega \to \mathbf{P}_f(Y)$ by

$$\mathbf{G}(\omega) = \begin{cases} \mathbf{F}(\omega), & \text{if } \omega \in \Omega \setminus N; \\ \{0\}, & \text{if } \omega \in N. \end{cases}$$

Therefore, $\mathbf{G}$ is a μ-measurable multifunction and, since X^* has the Radon–Nikodym property, Y^* is separable; so, by Theorem 3.1 of [23],

$$\mathbf{E}_{\mathbf{G}}^{\Sigma_n}(\omega) \xrightarrow{K.M.} \mathbf{E}_{\mathbf{G}}^{\Sigma_0}(\omega), \quad \mu \text{ a.e.}$$

Since $\mathbf{E}_{\mathbf{G}}^{\Sigma_n}$ is Σ_n-μ-measurable, there is a sequence $g_{n,m} : \Omega \to Y \subset X$ of Σ_n-μ-measurable functions so that $\mathbf{E}_{\mathbf{G}}^{\Sigma_n}(\omega) = \overline{\{g_{n,m}(\mu)\}}$ $\forall \omega \in \Omega$ and $\forall n \geq 0$.

Since $\mathbf{G}(\omega) = \mathbf{F}(\omega)$ (μ a.e),

$$\mathbf{E}_{\mathbf{G}}^{\Sigma_n}(\omega) = \mathbf{E}_{\mathbf{F}}^{\Sigma_n}(\omega)$$

(μ a.e.), $\forall n \geq 0$. Therefore,

$$\mathbf{E}_{\mathbf{F}}^{\Sigma_n}(\omega) \xrightarrow{K.M.} \mathbf{E}_{\mathbf{F}}^{\Sigma_0}(\omega),$$

μ a.e.

The same type of argument allows us to obtain the following result.

THEOREM 4.2. *If X^* has the Radon–Nikodym property, $\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(X)$ is a sequence of μ-measurable multifunctions $\mathbf{F}_n(\omega) \subset \mathbf{G}(\omega)$ (μ a.e.), with $\mathbf{G} : \Omega \to \mathbf{P}_{wkc}(X)$ integrably bounded, $\mathbf{F}_n(\omega) \xrightarrow{K.M.} \mathbf{F}(\omega)$ (μ a.e.), and $\Sigma_n \to \Sigma_0$ in $L^1_X(\mu)$; then*

$$\mathbf{S}^1_{\mathbf{E}_{\mathbf{F}}^{\Sigma_n}} \xrightarrow{K.M.} \mathbf{S}^1_{\mathbf{E}_{\mathbf{F}}^{\Sigma_0}}.$$

Proof. If X^* is separable, the result is Theorem 3.2 of [23]. In the general case, since each $\mathbf{F}_n$ is measurable, there exists an $N_n \in \mathbf{N}$ such that $\mu(N_n) = 0$ and $\bigcup_{\omega \in \Omega \setminus N_n} \mathbf{F}_n(\omega)$ is separable. If $N = \bigcup_{n=1}^{\infty} N_n$, then $N \in \Sigma$, $\mu(N) = 0$, and the closed subspace Y, generated by $\bigcup_{\omega \in \Omega \setminus N} \mathbf{F}(\omega)$, is separable. The remainder of the proof follows by using the same technique applied in the previous theorem. □

Sometimes it is possible to remove the Radon–Nikodym property in Theorem 4.2. To do that we need the following lemmas.

LEMMA 4.1. *A function $\mathbf{F} : \Omega \to \mathbf{P}_f(X)$ is μ-measurable if and only if there is an $N \in \Sigma$ so that $\mu(N) = 0$, $Y = [\bigcup_{\omega \in \Omega \setminus N} \mathbf{F}(\omega)]$ is separable, and $\overline{\mathbf{F}}(B) \in \Sigma$ for each closed ball B in the relative strong topology of Y.*

Proof. It is enough to prove the converse. Let A be a nonempty closed set of Y, set $H \subset A$ dense and countable, and $\varepsilon > 0$. If $U_\varepsilon = \bigcup_{x\in H} \overline{B}(x,\varepsilon)$, then

$$\mathbf{F}^-(U_\varepsilon) = \bigcup_{x\in H} \mathbf{F}^-(\overline{B}(x,\varepsilon)),$$

which implies that $\mathbf{F}^-(U_\varepsilon) \in \Sigma$. By setting $\varepsilon = \frac{1}{n}, n \in \mathbf{N}$, we get

$$A = \bigcap_{n=1}^{\infty} U_{\frac{1}{n}},$$

and therefore $\mathbf{F}^-(A) = \bigcap_{n=1}^{\infty} \mathbf{F}^-(U_{\frac{1}{n}}) \in \Sigma$. So, by Theorem 3.1, $\mathbf{F}$ is μ-measurable. $\square$

In the separable case, the preceding lemma is equivalent to (a) and (d) of Lemma 2.1 of [14]. The elementary proof given here is inspired by [9, p. 92].

LEMMA 4.2. *Let X be a Banach space and $\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(X)$ be a sequence of measurable multifunctions such that there is a weakly compact, convex, and separable subset of X, W such that $\mathbf{F}_n(\omega) \subset W, \forall w \in \Omega$ and $n \in \mathbf{N}$. Then $\mathbf{F}(\omega) = \overline{\lim}_n \mathbf{F}_n(\omega)$ is measurable, takes its values in $\mathbf{P}_{wkc}(X)$, and $\mathbf{F}(\omega) \subset W, \forall w \in \Omega$.*

Proof. If X is reflexive and separable, Proposition 4.3 of Hess [14] provides the proof with $\mathbf{G}(\omega) = W$. In the general case, the Davis–Fiegel–Johnson–Pelczynski factorization scheme [5] produces a separable reflexive Banach space R, a one-to-one bounded linear operator $\mathbf{J} : R \to X$, and a weakly compact convex subset K of R such that the restriction $\mathbf{J}|_K$ is a weak homeomorphism between K and W.

Now we consider the multifunctions

$$\mathbf{J}^{-1}\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(R).$$

$\mathbf{J}^{-1}\mathbf{F}_n(\omega) \subset K$ for each $\omega \in \Omega$ and $n \in \mathbf{N}$. Furthermore, given a closed ball B in Y, the closed subspace generated by K, we have

$$B \cap \mathbf{J}^{-1}\mathbf{F}_n(\omega) \neq \emptyset \iff (B \cap K) \cap \mathbf{J}^{-1}\mathbf{F}_n(\omega) \neq \emptyset,$$

which means that it is enough to consider closed balls in K.

Let B be a closed ball in K. Then $\mathbf{J}^{-1}\mathbf{F}_n^-(B) = \mathbf{F}_n^-(\mathbf{J}(B)) \in \Sigma$; hence Lemma 4.1 implies $\mathbf{J}^{-1}\mathbf{F}_n$ μ-measurable for each $n \in \mathbf{N}$. Now, applying Proposition 4.3 of [14], we get that

$$H = \overline{\lim}_{n\to\infty} \mathbf{J}^{-1}\mathbf{F}_n$$

is a μ-measurable multifunction with values in $\mathbf{P}_{wkc}(R)$ and $H(\omega) \subset K$, $\forall \omega \in \Omega$. Since $\mathbf{F}_n = \mathbf{J}\mathbf{J}^{-1}\mathbf{F}_n$ for each $n \in \mathbf{N}$ and $\mathbf{J}|_K$ is a weak homeomorphism between K and W, we get

$$\mathbf{F}^-(B) = \mathbf{J}H^-(B) = H^-(\mathbf{J}^{-1}B) \in \Sigma,$$

and Lemma 4.1 implies that $\mathbf{F}$ is μ-measurable. $\square$

THEOREM 4.3. *Let X be a Banach space and $\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(X)$ be a sequence of μ-measurable multifunctions. If there is a weakly compact, convex, separable subset W of X such that $\mathbf{F}_n(\omega) \subset W, \forall w \in \Omega$ and $\forall n \in \mathbf{N}$, and $\mathbf{F}_n \xrightarrow{K.M.} \mathbf{F}$, then*

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{F}_n}} \xrightarrow{K.M.} \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{F}}}$$

whenever Σ_n *is a sequence of sub-*σ*-fields that converges to* Σ_0 *in* $L^1_X(\mu)$.

Proof. Since W is separable, the reflexive Banach space R tailored through the Davis–Fiegel–Johnson–Pelczynski factorization scheme is separable. Let K be a weakly compact subset of R so that $\mathbf{J}|_K : K \to W$ is a weak homeomorphism. From Lemma 4.2, $\mathbf{J}^{-1}\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(R)$ is a sequence of μ-measurable multifunctions, and so, there is another μ-measurable multifunction $\mathbf{H} : \Omega \to \mathbf{P}_{wkc}(R)$ so that $\mathbf{H}(\omega) \subset K$, $\forall \omega \in \Omega$, where $H = \overline{\lim}\mathbf{J}^{-1}\mathbf{F}_n$. By taking $G \equiv K$, we can apply Theorem 4.2 of [21] in order to get

$$\omega - \overline{\lim}\,\mathbf{S}^1_{\mathbf{J}^{-1}\mathbf{F}_n} \subset \mathbf{S}^1_{\mathbf{H}}.$$

Being R-reflexive and separable, as is its dual, and from the proof of Theorem 3.2 of [23],

$$\omega - \overline{\lim}\,\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}} \subset \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{H}}}.$$

The same proof also shows that

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{F}}} \subset s - \underline{\lim}\,\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}.$$

Now, proceeding as in Diestel [6], we define a one-to-one bounded linear operator $\tilde{\mathbf{J}} : L^1_R(\mu) \to L^1_X(\mu)$ by $\tilde{\mathbf{J}}(f) = \mathbf{J} \circ f$ (it is well defined by Hille's theorem [8, p. 47]). We claim that

$$\omega - \overline{\lim}\,\tilde{\mathbf{J}}(\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}) \subset \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{J}(\mathbf{H})}}.$$

In fact, for each

$$f_n \in \tilde{\mathbf{J}}(\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}),$$

there is a $g_n \in \mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}$ such that $f_n = \mathbf{J}g_n$; so if

$$f \in \omega - \overline{\lim}\,\tilde{\mathbf{J}}(\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}),$$

there is an increasing sequence of positive integer numbers $\{n_k\}$ such that

$$g_{n_k} \in \mathbf{S}^1_{\mathbf{E}^{\Sigma_{n_k}}_{\mathbf{J}^{-1}\mathbf{F}_n{}_{n_k}}}$$

and $\mathbf{J}g_{n_k}$ converges to f in the weak topology of $L^1_X(\mu)$. Since

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}} \subset \{f : \Omega \to R : f \,\text{is a measurable selector of}\, \mathbf{G} \equiv \mathbf{K}\},$$

Theorem 3.2 implies that there is a μ-measurable function $g : \Omega \to K$ such that $g_{n_{k_l}}$ converges to g in the weak topology of $L^1_R(\mu)$ for some subsequence $g_{n_{k_l}}$ of g_{n_k}; so $g \in \mathbf{H}$. Hence, given $x^* \in (L^1_X(\mu))^*$, we have

$$x^* f_{n_{k_l}} = x^* \tilde{\mathbf{J}} g_{n_{k_l}} \xrightarrow{w} x^* \tilde{\mathbf{J}} g,$$

and therefore $f = \tilde{\mathbf{J}}g$ (μ a.e.), and this implies $f \in \mathbf{J}(\mathbf{H})$. But

$$\tilde{\mathbf{J}}(\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}}) = \mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{F}_n}}\ \forall n \quad \text{and} \quad \tilde{\mathbf{J}}(\mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{H}}) = \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{J}(\mathbf{H})}}.$$

This fact, together with the previous inclusion, gives us

$$\omega - \overline{\lim}\, \mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{\mathbf{J}^{-1}\mathbf{F}_n}} \subset \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{\mathbf{J}(\mathbf{H})}}.$$

It remains to prove that $\mathbf{J}(\mathbf{H}) = \mathbf{F}$, since then the conclusion follows from the obtained inclusions. By hypothesis $\mathbf{F}_n \xrightarrow{K.M.} \mathbf{F}$. Hence $\omega - \overline{\lim}\, \mathbf{F}_n = \mathbf{F}$. $\mathbf{J}$ being a weak homomorphism between K and W, we get $\mathbf{H} = \mathbf{J}^{-1}\mathbf{F}$. This ends the proof. □

Our next result brings to this context an application from operator theory.

DEFINITION 4.1. *Given X, Y Banach spaces, a bounded linear operator $T : X \to Y$ is called an Asplund operator if there are a Banach space Z and bounded linear operators $T_1 : X \to Z$, $T_2 : Z \to Y$ such that Z^* has the Radon–Nikodym property and $T = T_2 \circ T_1$.*

This notion was introduced and studied by Stegall. It has shown to be useful in convex analysis [26] and control theory [2]. Here is a small contribution to the theory of measurable multifunctions.

THEOREM 4.4. *Let (Ω, Σ, μ) be a finite complete measure space, X, Y Banach spaces, and $T : X \to Y$ an Asplund operator. If $\mathbf{F}, \mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(X)$, $n \in \mathbf{N}$, are measurable multifunctions, if there is an integrably bounded multifunction $G : \Omega \to \mathbf{P}_{wkc}(X)$, such that $\mathbf{F}_n(\omega) \subset \mathbf{G}(\omega)$ (μ a.e.) $\forall n, \in \mathbf{N}$, and if $\{\Sigma_n\}$ is a sequence of sub-σ-fields that converges to Σ_0 in $L^1_X(\mu)$, then*

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{T(\mathbf{F}_n)}} \xrightarrow{K.M.} \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{T(\mathbf{F})}}.$$

Proof. Let Z be a Banach space such that its dual Z^* has the Radon–Nikodym property and $T_1 : X \to Z$, $T_2 : Z \to Y$ bounded operators so that $T = T_2 \circ T_1$. If $W = \overline{T_1(X)}$, then Σ_n converges to Σ_0 in $L^1_W(\mu)$, and W^* has the Radon–Nikodym property. Furthermore, $T = T_3 \circ T_1$ where $T_3 = T_2|_W$. For each $n \in \mathbf{N}$, $T_1\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(X)$ is measurable and is contained in $T_1\mathbf{G} : \Omega \to \mathbf{P}_{wkc}(X)$, which is integrably bounded. Furthermore, $T_1\mathbf{F}_n \xrightarrow{K.M.} T_1\mathbf{F}$. So by Theorem 4.3,

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{T_1(\mathbf{F}_n)}} \xrightarrow{K.M.} \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{T_1(\mathbf{F})}}.$$

Therefore,

$$\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{T(\mathbf{F}_n)}} = \mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{T_3 \circ T_1(\mathbf{F}_n)}} = T_3(\mathbf{S}^1_{\mathbf{E}^{\Sigma_n}_{T_1(\mathbf{F}_n)}}) \xrightarrow{K.M.} T_2(\mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{T_1(\mathbf{F})}}) = \mathbf{S}^1_{\mathbf{E}^{\Sigma_0}_{T(\mathbf{F})}}. \qquad □$$

5. An application in control theory. In this section we give some applications of the Kuratowski–Mosco convergence into optimal control theory allowing the controls to be variable in [0, T]. Therefore, in this section, μ is the Lebesgue measure in [0,T].

Given two Banach spaces U, X and a bounded linear operator $B : U \to X$, we consider the following linear systems in X:

$$\mathbf{S}_n = \begin{cases} \dot{x}(t) = Ax(t) + Bf^{(n)}(t), \\ x(0) = x_0; \end{cases}$$

$$\mathbf{S} = \begin{cases} \dot{x}(t) = Ax(t) + Bf(t), \\ x(0) = x_0, \end{cases}$$

where $f^{(n)}, f$ are μ-measurable selectors of $\mathbf{F}_n : [0,T] \to \mathbf{P}_{wkc}(U)$, $\mathbf{F} : [0,T] \to \mathbf{P}_{wkc}(U)$, where $\mathbf{F}_n$, $\mathbf{F}$ are strongly μ-measurable multifunctions and A generates a strongly continuous semigroup of bounded linear operators $\{S_t\}_{t\geq 0}$. The set $K(t)$, defined as

$$K(T) = \left\{ \mathbf{S}_T x_0 + \int_0^T \mathbf{S}_{T-s} Bf(s) ds, f \in \mathbf{S}^1_{\mathbf{F}} \right\},$$

is called the set of attainable points for the system S in the time T. Analogously, we define the set $K_n(T)$ as the set of attainable points for the system S_n in the time T (where we consider $f \in \mathbf{S}^1_{\mathbf{F}_n}$ instead).

The system S (or S_n) is controllable in time T if $0 \in K(T)$ (respectively, $0 \in K_n(T)$). Controllability criteria for these classes of systems with $\mathbf{F}$ constant can be found in [2], [3], and [24].

The aim of this section is the following result.

Theorem 5.1. *Let $\mathbf{S}_n$, $\mathbf{S}$, $\mathbf{F}_n$, and $\mathbf{F}$ be as above. If $\mathbf{F}_n \xrightarrow{K.M.} \mathbf{F}$ and there is a $\mathbf{G} : [0,T] \to \mathbf{P}_{wkc}(X)$, integrably bounded such that for each $n \in \mathbf{N}$, $\mathbf{F}_n(\omega) \subset \mathbf{G}(\omega)$, μ a.e., then $K_n(T) \xrightarrow{K.M.} K(T)$.*

Proof. Since $\mathbf{F}_n \xrightarrow{K.M.} \mathbf{F}$, then for each $t \in [0,T]$, $\mathbf{S}_t B\mathbf{F}_n \xrightarrow{K.M.} \mathbf{S}_t B\mathbf{F}$. Furthermore, $\mathbf{S}_t B\mathbf{F}_n : \Omega \to \mathbf{P}_{wkc}(U)$ and are strongly μ-measurable with $\mathbf{S}_t B\mathbf{F}_n(\omega) \subset \mathbf{S}_t B\mathbf{G}(t)$ (μ a.e.) with $\mathbf{S}_t B\mathbf{G}$ convex weakly compact-valued and integrably bounded. Now, applying Theorem 3.4 of [21], we get

$$\int_0^T \mathbf{S}_{t-s} B\mathbf{F}_n(s) ds \xrightarrow{K.M.} cl \int_0^T \mathbf{S}_{t-s} B\mathbf{F}(s) ds,$$

which implies the conclusion because $K_n(T)$ and $K(T)$ are merely translations of $\int_0^T \mathbf{S}_{t-s} B\mathbf{F}_n(s) ds$ and $\int_0^T \mathbf{S}_{t-s} B\mathbf{F}(s) ds$, respectively. □

Corollary 5.1. *Under the same hypothesis as in Theorem* 5.1, *if $\mathbf{S}_n$ is controllable at time T and $\mathbf{F}_n \xrightarrow{K.M.} \mathbf{F}$, then $\mathbf{S}$ is controllable at time T.*

Proof. The proof follows immediately from the definitions and Theorem 5.1. □

6. Some applications in differential inclusions. Now we give a set of results related to the study of compactness of the solution of certain linear systems of differential inclusions.

In this entire section, $\Omega = [a,b] \subset \mathbf{R}$, Σ is the Borel σ-field on $[a,b]$ and μ is the Lebesgue measure on $[a,b]$.

We start with an easy result that has some interesting consequences.

Proposition 6.1. *Let X be a Banach space and $K \subset L^1_X(\mu)$ be uniformly integrable. The set*

$$A = \left\{ x : [a,b] \to X : x(t) = \int_a^t f d\mu; f \in K \right\}$$

is relatively compact in the uniform topology of $C([a,b];X)$ if and only if, for each $t \in [a,b]$, $\{x(t)\}_{x\in A}$ is relatively compact in the norm topology of X.

Proof. The boundedness and equicontinuity of A follows from the uniform integrability of K, and then the conclusion follows from the Arzéla–Ascoli theorem. □

The following result is due to Fattorini [12, Thm. 3.1] for the case $a = 0$. We present a simpler proof even though we use the same ideas as Fattorini.

THEOREM 6.1. *Let K be a uniformly integrable subset of $L^1_X(\mu)$ and $\mathbf{S}_t : X \to X$ a strongly continuous semigroup that is compact for each $t0$. Then, for $0 \leq a < b < \infty$, the set*

$$A = \left\{ x : [a,b] \to X : x(t) = \int_a^t \mathbf{S}_{t-s} f d\mu; f \in K \right\}$$

is relatively compact in the uniform topology of $C([a,b];X)$.

Proof. Let us suppose first that $a = 0$. Since K is uniformly integrable in $L^1_X(\mu)$ and $\mathbf{S}_{(\cdot)}$ is uniformly bounded on [0,b], we get that $\{\mathbf{S}_{(\cdot)} f\}_{f \in K}$ is uniformly integrable in $L^1_X(\mu)$. Let $t \in [0,b]$ and δ_n be a sequence of real numbers such that $\delta_n \to t$. Notice that for each $n \in \mathbf{N}$ and $f \in L^1_X(\mu)$,

$$\begin{aligned}\int_0^t \mathbf{S}_{t-s} f(s) d\mu(s) &= \int_0^{t-\delta_n} \mathbf{S}_{t-s} f(s) d\mu(s) + \int_{t-\delta_n}^t \mathbf{S}_{t-s} f(s) d\mu(s) \\ &= \mathbf{S}_{\delta_n} \left(\int_0^{t-\delta_n} \mathbf{S}_{t-\delta_n - s} f(s) d\mu(s) \right) + \int_{t-\delta_n}^t \mathbf{S}_{t-s} f(s) d\mu(s).\end{aligned}$$

The set $\{\int_0^t \mathbf{S}_{t-s} f(s) d\mu(s)\}_{f \in K}$ is bounded in X, and by hypothesis, $\mathbf{S}_{\delta_n}$is a compact operator; therefore, $\{\int_0^{t-\delta_n} \mathbf{S}_{t-s} f(s) d\mu(s)\}_{f \in K}$ is relatively compact in X. Thus, given a sequence $\{f_k\}$ in K, we can find, by the diagonal process, a subsequence $\{f_{k_l}\}$ of $\{f_k\}$ such that

$$\left\{ \int_0^{t-\delta_n} \mathbf{S}_{t-s} f_{k_l}(s) d\mu(s) \right\}$$

is convergent in X. Now, using the uniform integrability of K, we can see that $K - K = \{f - g : f, g \in K\}$ is also uniformly integrable. Therefore, given $\varepsilon > 0$, there are $l_0, n_0 \in \mathbf{N}$ such that for each $l > l_0$,

$$\begin{aligned}\left\| \int_0^t \mathbf{S}_{t-s}(\{f_{k_l}(s) - f_{k_{l_0}}(s)) d\mu(s) \right\| &\leq \left\| \int_0^{t-\delta_{n_0}} \mathbf{S}_{t-s}(f_{k_l}(s) - f_{k_{l_0}}(s)) d\mu(s) \right\| \\ &\quad + \left\| \int_{t-\delta_{n_0}}^t \mathbf{S}_{t-s}(f_{k_l}(s) - f_{k_{l_0}}(s)) d\mu(s) \right\| < \varepsilon,\end{aligned}$$

and now the conclusion follows from the Arzéla–Ascoli theorem.

In the general case, $a \neq 0$, we identify $L^1_X([a,b],\mu)$ with $L^1_X([0,b-a],\mu)$ using the identification $f \longleftrightarrow g$ if and only if $g(s) = f(x+a)$. Let K be uniformly integrable in $L^1_X([a,b],\mu)$ and K' its image under the identification. It is easy to see that K' is uniformly integrable in $L^1_X([0,b-a],\mu)$, and therefore, applying Proposition 6.1 together with the first part of the proof, we conclude that

$$\left\{ \int_0^t \mathbf{S}_{t-a-s} g(s) d\mu(s) \right\}_{g \in K'}$$

is relatively compact in X. This means that

$$\begin{aligned}\left\{\mathbf{S}_a\left(\int_0^{t-a}\mathbf{S}_{t-a-s}g(s)d\mu(s)\right)\right\}_{g\in K'} &= \left\{\left(\int_0^{t-a}\mathbf{S}_{t-s}g(s)d\mu(s)\right)\right\}_{g\in K'}\\ &= \left\{\left(\int_a^{t}\mathbf{S}_{t-s}g(s-a)d\mu(s)\right)\right\}_{g\in K'}\\ &= \left\{\left(\int_a^{t}\mathbf{S}_{t-s}f(s)d\mu(s)\right)\right\}_{f\in K}\end{aligned}$$

is relatively compact in K. Now, by Proposition 6.1, we conclude the proof. □

Now we state the main result of this section.

THEOREM 6.2. *Let* $\mathbf{F}\to\mathbf{P}_{wkc}(X)$ *be an integrable bounded multifunction,* $\{\mathbf{S}_t\}_{t\geq 0}$ *a strongly continuous semigroup of operators in* X*, and*

$$A=\left\{x\in C([a,b];X) : x(t)=\int_a^t\mathbf{S}_{t-s}f ds\right\}_{f\in\mathbf{S}^1_{\mathbf{F}}}.$$

Then,

(i) *if* $\mathbf{S}_t$ *is compact for each* $t>0$*,* A *is compact in* $C([a,b];X)$*;*

(ii) *if there is a compact set* $K\subset X$ *such that* $\mathbf{F}(t)\subset K$*,* $\forall t\in[a,b]$*,* A *is compact in* $C([a,b];X)$*.*

Proof. (i) From Proposition 6.1, we know that A is relatively weakly compact in $C([a,b];X)$. Let x_n, x be in $C([a,b];X)$, such that $||x_n-x||_\infty\to 0$. Clearly, for each $t\in[a,b]$, $x_n(t)$ converges weakly to $x(t)$. On the other hand, there is a sequence $\{f_n\}\subset\mathbf{S}^1_{\mathbf{F}}$ such that, for each $n\in\mathbf{N}$,

$$x_n(t)=\int_a^t\mathbf{S}_{t-s}f_n(s)d\mu(s),$$

since Theorem 3.2 ensures that $\{\mathbf{S}_{t-(\cdot)}f_n(\cdot)\}_{f\in\mathbf{S}^1_{\mathbf{F}}}$ is weakly compact, there are $g\in\mathbf{S}^1_{\mathbf{F}}$ and a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ such that $\mathbf{S}_{t-(\cdot)}f_{n_k}(\cdot)$ converges to $\mathbf{S}_{t-(\cdot)}g(\cdot)$ in the weak topology of $L^1_X(\mu)$. This implies that for each $t\in[a,b]$, $x_{n_k}(t)$ converges weakly to $y(t)=\int_a^t\mathbf{S}_{t-s}g(s)d\mu(s)$. Hence $x\equiv y$.

(ii) Without loss of generality, we can suppose that K is convex and separable. By Corollary 8 of [8, p. 48], we have that for each $t\in(a,b]$, $\frac{1}{t}\int_a^t f(s)ds\in K$. Set $G\equiv K$, and let $B=\{x : x(t)=\int_a^t f(s)ds\}_{f\in\mathbf{S}^1_{\mathbf{G}}}$. Then $\{x(t)\}_{x\in B}$ is relatively compact in X, and thus B is relatively compact in $C([a,b];X)$.

Let $\{f_n\}_{n=1}^\infty$ be a sequence of simple functions in $\mathbf{S}^1_{\mathbf{F}}$, write the image of f_n as

$$\mathrm{Im}f_n=\{x_{1(n)},x_{1(n)},\dots,x_{k(n)}\},$$

and define

$$A_{n_i}=\{t\in[a,b] : f_n(t)=x_i(n),\ i=1,2,\dots,k(n)\}.$$

For each $n\in\mathbf{N}$ we have that $\{A_{n_i}\}_{i=1}^{k(n)}$ is a finite partition of $[a,b]$. Since $\{\int_a^t f_n(s)d\mu(s)\}_{n\in\mathbf{N}}$ is relatively compact, then given $\varepsilon>0$ and $n_0\in\mathbf{N}$, there is a subsequence $\{f_{n_k}\}_{k,n=1}^\infty$ of $\{f_n\}_{n=1}^\infty$ such that

$$\left\|\int_a^t(f_{n_0}(s)-f_m(s))d\mu(s)\right\|<\varepsilon\quad\forall m>n_0.$$

By means of elementary operations with the sets A_{n_k}, we construct a finite partition (for each $m > n_0$) $\{B_j\}_{j=1}^{l}$ of $[a,b]$, such that each B_j is measurable and $f_k|_{B_j}$ is constant for each k and j ($1 \leq j \leq l$, $1 \leq k \leq m$). Therefore $\{\bigcup_{k,j}^{l,m} f_k(B_j)\} = \{x_1, \ldots, x_r(m)\}$ is a finite subset of X. For each $n \geq n_0$, denote by X_n the subspace generated by $\{x_1, \ldots, x_r(n)\}$ and consider the functions $\mathbf{S}_{(\cdot)}|_{X_n}$. Since each X_n is finite dimensional and $\mathbf{S}_{(\cdot)}$ is strongly continuous, we get that $\mathbf{S}_{(\cdot)}|_{X_n}$ is continuous in the uniform topology of operators on $L(X_n, X)$ and is therefore measurable. Consequently, there is a $k \in \mathbf{N}$ such that $\mathbf{S}^k : [a,b] \to L(X_n, X)$ is a simple function and

$$||\mathbf{S}^k_{(\cdot)} - \mathbf{S}|_{X_m}(\cdot)|| < \varepsilon.$$

By using Hille's theorem [8, p. 43] we have that for each $T \in L(X_m, X)$,

$$\left\|\int_a^t T(f_{n_0}(s) - f_m(s))d\mu(s)\right\| < ||T||\varepsilon$$

for $m > n_0$, and from this we conclude that $\{\int_a^b \mathbf{S}^k(f_n)ds\}_{n=1}^{\infty}$ is relatively compact in $C([a,b], X)$. By this procedure we see that

$$\begin{aligned}\left\|\int_a^t \mathbf{S}_s(f_{n_0}(s) - f_m(s))d\mu(s)\right\| &\leq \left\|\int_a^t \mathbf{S}_s(f_{n_0}(s)) - \mathbf{S}^k(f_{n_0}(s))d\mu(s)\right\| \\ &+ \left\|\int_a^t \mathbf{S}^k(f_m(s)) - \mathbf{S}^k(f_{n_0}(s))d\mu(s)\right\| \\ &+ \left\|\int_a^t \mathbf{S}_s(f_m(s) - \mathbf{S}^k(f_m(s))d\mu(s)\right\| \\ &\leq 2\varepsilon^2 M + \varepsilon||\mathbf{S}^k||,\end{aligned}$$

where $M = \sup_{x \in K}||x||$. This implies that $\{\int_a^t \mathbf{S}_{t-s}f(s)ds : \text{f simple}\}_{f \in \mathbf{S}^1_{\mathbf{G}}}$ is relatively weakly compact in X. Now using the boundedness of $\mathbf{S}_{(\cdot)}$ on $[a,b]$ and the density of simple functions on $L^1_X(\mu)$, we conclude that

$$B = \left\{x : [a,b] \to X : x(t) = \int_a^t f(t)dt\right\}_{f \in \mathbf{S}^1_{\mathbf{G}}}$$

is relatively compact in $C([a,b]; X)$, and since

$$A = \left\{x : [a,b] \to X : x(t) = \int_a^t f(t)dt\right\}_{f \in \mathbf{S}^1_{\mathbf{F}}} \subset B,$$

then A is relatively compact in $C([a,b]; X)$.

To see that A is compact, we note that for each $s \in [a,b]$, $\mathbf{S}_s\mathbf{F}(s)$ is a compact, convex set-valued multifunction, and using Theorem 3.2 as in (i), we conclude the proof. □

The hypothesis "$\mathbf{F}(s)$ contained in a compact set K for each $s \in [a,b]$" cannot be removed, as the following example illustrates.

Example. Let X be an infinite-dimensional Banch space and K a bounded, separable and convex subset of X which is not compact. Define $\mathbf{F}(s) \equiv K, \forall s \in [0,b]$, and

$\mathbf{S}_s \equiv I$, the identity in $X, \forall s \geq 0$. If $\{x_n\}_{n=1}^{\infty}$ is a sequence in K with no convergent subsequences, then for each $t \in [0, b]$, we have

$$\left\{\int_0^t \mathbf{S}_{t-s} x_k ds\right\}_{k\in\mathbf{N}} = \{tx_k\}_{k\in\mathbf{N}},$$

which is relatively compact if and only if $t = 0$. □

We end this section by illustrating how these results can be useful in the qualitative study of the trajectories of differential inclusions in Banach spaces. To do that, we start by outlining the path developed by Frankowska [13] but remove the hypothesis of *separability and reflexibility*. In that reference the following differential inclusion is considered:

$$\mathbf{S} = \begin{cases} \dot{x}(t) \in Ax(t) + \mathbf{F}(t, x), \\ x(t_0) = x_0, \end{cases}$$

where A generates a strongly continuous semigroup $\{\mathbf{S}_t\}_{t\geq 0}$ and $\mathbf{F} : [to, T] \times X \to \mathbf{P}(X)\backslash\{\emptyset\}$ is a multifunction.

In what follows, we need the following definitions.

Let $\mathbf{F} : [to, T] \times X \to \mathbf{P}_f(X)$, $t_0 \in [0, T]$. A function $x \in C([t_0, T]; X)$ is called a *mild trajectory* for the given differential inclusion if there is an $f \in L^1([t_0, T], X)$ such that

$$f(t) \in \mathbf{F}(t, x(t)), \quad \mu \text{ a.e.},$$

and

$$x(t) = \mathbf{S}_{t-t_0} x_0 + \int_{t_0}^t \mathbf{S}_{t-s} f(s) ds.$$

A function $\varphi : X \to \mathbf{P}(X)\backslash\{\emptyset\}$ is called an *L-Lipschitz* in $K \subset X$ if, for $x, y \in X$,

$$\varphi(x) \subset \varphi(y) + L||x - y||B,$$

where B is the closed unit ball of X.

With these hypotheses in hand, the following result is proposed [13, Thm. 2.7].

Let X be a reflexive (separable) Banach space and $\mathbf{F} : [t_o, T] \times X \to X$ be a multivalued function with closed and convex values. Suppose that there is a $\mathbf{k} \in L^1_{(t_0;T)}$ so that for almost every $t \in [t_0, T]$, $\mathbf{F}(t, \cdot)$ is $\mathbf{k}(t)$-Lipschitz, and for each $x \in X$, $\mathbf{F}(t, x) \subset \mathbf{k}(t)B$. If at least one of the following conditions is satisfied,

(i) the semigroup $\mathbf{S}_{(\cdot)}$ is compact;

(ii) the semigroup $\mathbf{S}_{(\cdot)}$ is uniformly continuous;

(iii) there is a compact set $K \subset X$ such that for each $(t, x) \in [t_0, T] \times X$, $\mathbf{F}(t, x) \subset \mathbf{K}$; then for each $x \in X$, the set $\Gamma_{[t_0,T]}(\xi)$ of the mild solutions of the differential inclusion with initial values $x(t_0) = \xi$ is a compact subset of $C([t_0, T]; X)$.

According to Theorem 6.2, it is clear that statements (i) and (iii) are true without any restriction on the Banach space X, while the statement (ii) fails to be true in any infinite-dimensional Banach space, as is shown in the previous example. This is so because in the process of the proof given in [13], the relative compactness of $\{x(t)\}_{x\in S(t_0,T,X)}$ is not checked before applying the Arzéla–Ascoli theorem for infinite-dimensional Banach spaces. In that sense, Theorem 6.2 not only generalizes Theorem 2.7 (i), (ii) of [13] but also gives a correct proof of these results.

However, that paper is still interesting, since it states the problem to be solved and offers other interesting results about the subject.

REFERENCES

[1] Z. Artstein, *Weak convergence of set valued functions and Control*, SIAM J. Control Optim., 13 (1975), pp. 865–878.

[2] D. Bárcenas and J. Diestel, *Constrained controllability in non-reflexive Banach spaces*, Quaestiones Matematicae, 18 (1995), pp. 185–198.

[3] D. Bárcenas and H. Leiva, *Controlabilidad con restricciones en Espacios de Banach*, Acta Cientifica Venezolana (Matemáticas), 40 (1989), pp. 181–185.

[4] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.

[5] W. J. Davis, T. Fiegel, W. B. Johnson, and A. Pelczynski, *Factoring weakly compact operators*, J. Functional Anal., 17 (1974), pp. 311–327.

[6] J. Diestel, *Remarks on weak compactness in* $L^1(\mu, X)$, Glasgow Math. J., 18 (1977), pp. 87–92.

[7] J. Diestel, J. Ruess, and W. Schachermayer, *On weak compactness in* $L^1(\mu, X)$, Proc. Amer. Math. Soc., 118 (1993), pp. 443–453.

[8] J. Diestel and J. J. Uhl, *Vector Measures*, Amer. Math. Soc. Surveys, Vol.15, AMS, Providence, RI, 1977.

[9] N. Dinculeanu, *Vector Measures*, Pergamon Press, New York, 1967.

[10] D. van Dulst, *The Geometry of Banach spaces with the Radon–Nikodym property*, Rend. Circ. Mat. Palermo, 2 (1985), 81 pp.

[11] L. Egghe, *Stopping Time Techniques for Analysts and Probabilists*, London Math. Soc. Lecture Notes 100, Cambridge University Press, Cambridge, UK, 1984.

[12] H. O. Fattorini, *Relaxation in Semilinear infinite dimensional control systems*, Differential Equations, Dynamical Systems and Control Science, Marcel Dekker, New York, 1993, pp. 505–522.

[13] H. Frankowska, *A priori Estimates for Operational Differential Inclusions*, J. Differential Equations, 84 (1990), pp. 100–128.

[14] C. Hess, *Measurability and integrability of the weak upper limit of a sequence of multifunctions*, J. Math. Anal. Appl., 153 (1990), pp. 206–249.

[15] F. Hiai and H. Umegaki, *Integrals, conditional expectations and martigales of multivalued functions*, J. Mult. Anal., 7 (1977), pp. 149–182.

[16] R. C. James, *Weakly compact sets*, Trans. Amer. Math. Soc., 113 (1964), pp. 129–140.

[17] U. Mosco, *Convergence of convex sets of some variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

[18] N. Papageorgiou, *On the theory of Banach spaces valued multifunctions* I*, integration and conditional expectation*, J. Mult. Anal., 17 (1985), pp. 185–206.

[19] N. Papageorgiou, *Representation of set valued operators*, Trans. Amer. Math. Soc., 292 (1985), pp. 557–572.

[20] N. Papageorgiou, *On the efficiency and optimality of allocations* II, SIAM J. Control Optim., 24 (1986), pp. 452–479.

[21] N. Papageorgiou, *Convergence Theorems for Banach space valued multifunctions*, Internat. J. Math. & Math. Sci., 10 (1987), pp. 433–442.

[22] N. Papageorgiou, *Convergence and representation theorems for set valued random processes*, J. Math. Anal. Appl., 150 (1990), pp. 129–145.

[23] N. Papageorgiou, *Convergence theorems for set valued conditional expectations*, Comm. Math. Univ. Carolinae, 34 (1993), pp. 97–104.

[24] G. Peichl and W. Schappcher, *Constrained controllability in Banach spaces*, SIAM J. Control Optim., 24 (1986), pp. 1261–1275.

[25] C. Stegall, *The Radon–Nikodym property in conjugate Banach spaces*, Trans. Amer. Math. Soc., 206 (1975), pp. 213–223.

[26] C. Stegall, *The Radon–Nikodym property in conjugate Banach spaces* II, Trans. Amer. Math. Soc., 264 (1981), pp. 507–519.

[27] A.Ülger, *Weak compactness in Banach spaces*, Proc. Amer. Math. Soc., 113 (1991), pp. 143–149.

SIAM J. MATH. ANAL.
Vol. 28, No. 5, pp. 1227–1247, September 1997

014

MONODROMY GROUPS OF SYSTEMS OF TOTAL DIFFERENTIAL EQUATIONS OF TWO VARIABLES*

TOSHIAKI YOKOYAMA†

Abstract. Fuchsian systems of ordinary differential equations that are irreducible and free from accessory parameters are classified into eight classes. Canonical systems of these classes were determined by Haraoka [SIAM J. Math. Anal., 25 (1994), pp. 1203–1226]. Among them the canonical systems of exactly four classes are extended to the completely integrable systems of total differential equations of two variables that are introduced by Yokoyama [Funkcial. Ekvac., 35 (1992), pp. 65–93]. This paper presents monodromy groups of the four systems of total differential equations. The theory for computing monodromy groups of the systems developed by Yokoyama is applied.

Key words. monodromy group, system of total differential equations, accessory parameter

AMS subject classifications. 33E30, 33C65, 34A20

PII. S0036141096296875

Introduction. In the theory of differential equations in the complex domain the Gauss hypergeometric differential equation

$$x(1-x)\frac{d^2y}{dx^2}+\{\gamma-(\alpha+\beta+1)x\}\frac{dy}{dx}-\alpha\beta y=0 \tag{0.1}$$

is one of the most important and interesting equations, and there are many investigations and generalizations of it. The Okubo theory is one of them. Okubo recognized a system of linear ordinary differential equations of the form

$$(xI_n-T)\frac{dY}{dx}=AY, \tag{0.2}$$

where Y is an unknown n-dimensional column vector, A is an $n\times n$ matrix, T is an $n\times n$ diagonal matrix, and I_n is the identity matrix of rank n, as a generalization of the Gauss equation (0.1), and developed a global theory of the system in [4]. His theory, which seems to be an extension of Riemann's P-function method for computing a monodromy group of the Gauss equation, consists of the following three parts: (i) reduction of a single Fuchsian differential equation to the system, (ii) definition of systems free from accessory parameters, and (iii) an algorithm for computing monodromy groups for such systems. He asserted that if the system (0.2) is free from accessory parameters, then we can determine its monodromy group explicitly up to a diagonal transformation. After Okubo, the present author [10] classified the set of irreducible systems free from accessory parameters. There are eight classes of such systems. Haraoka [2] determined canonical systems of these classes: systems (I), (I*), (II), (II*), (III), (III*), (IV), and (IV*). System (I) is known to be transformed into the generalized hypergeometric equation (see [4], [5]). System (I*) is known to be transformed into the Pochhammer equation (see [1], [8]). Systems (II), (II*), (III), (III*), (IV), and (IV*) are new. In [3] Haraoka also computed monodromy groups of these six systems by following the Okubo theory. Monodromy groups of system (II*)

*Received by the editors January 2, 1996; accepted for publication April 17, 1996.
http://www.siam.org/journals/sima/28-5/29687.html

†Department of Mathematics, Chiba Institute of Technology, Narashino, Chiba 275, Japan (yokoyama@cc.it-chiba.ac.jp).

of rank 4 and system (II) are also computed by Sasai [6] and Sasai and Tsuchiya [7], respectively.

In [9] the present author obtained an extension of the system (0.2) and investigated its monodromy group. If the coefficient matrix A satisfies the relation

$$(A-\rho_1 I_n)(A-\rho_2 I_n)=O, \tag{0.3}$$

where ρ_1 and ρ_2 are complex numbers, then we can extend the system (0.2) to a completely integrable system of total differential equations of two variables of the form

$$dZ=\Big\{(xI_n-T)^{-1}Adx+(A-(\rho_1+\rho_2)I_n)(yI_n-T)^{-1}dy-A\frac{d(x-y)}{x-y}\Big\}Z, \tag{0.4}$$

where Z is an unknown n-dimensional column vector. Among the eight systems we can extend systems (I*), (II*), (III*), and (IV*) to obtain systems of total differential equations, which we call systems (I**), (II**), (III**), and (IV**), respectively. It is shown in [9] that the systems of partial differential equations for the Appell functions F_1 and F_2 are reduced to system (I**) of rank 3 and system (II**) of rank 4, respectively.

The purpose of this paper is to compute monodromy groups of systems (I**), (II**), (III**), and (IV**) by following the theory developed in [9]. In section 1 we review the theory of the system (0.4). In section 2 we investigate system (I**). We give generators of a monodromy group of system (I**) in Theorem 1 in section 2.1 and prove the theorem in section 2.2. We deal with the other systems in section 3. We give generators of monodromy groups of systems (II**), (III**), and (IV**) in Theorems 2, 3, and 4 in sections 3.1, 3.2, and 3.3, respectively, and we prove these theorems in section 3.4. We use the following notation throughout this paper:

$\mathbf{Z}_{<0}$: the set of negative integers.
I_k : the identity matrix of rank k, for $k\in\mathbf{N}$.
O : zero matrix of an appropriate size.
$e(\alpha):=\exp(2\pi\sqrt{-1}\,\alpha)$, for $\alpha\in\mathbf{C}$.

1. Monodromy group of the system (0.4).

We give a short review of the theory developed in [9]. In the system (0.4) we assume that

$$T=\begin{pmatrix} t_1I_{n_1} & & & \\ & t_2I_{n_2} & & \\ & & \ddots & \\ & & & t_pI_{n_p}\end{pmatrix}\quad (t_i\neq t_j\ (i\neq j),\ n_1+n_2+\cdots+n_p=n),$$

$$A=\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p}\\ A_{21} & A_{22} & \cdots & A_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ A_{p1} & A_{p2} & \cdots & A_{pp}\end{pmatrix}\sim\begin{pmatrix}\rho_1 I_{m_1} & \\ & \rho_2 I_{m_2}\end{pmatrix}\quad (m_1+m_2=n),$$

where A_{ij} is an $n_i\times n_j$ matrix and

$$A_{ii}=\begin{pmatrix}\lambda_{i,1} & & & \\ & \lambda_{i,2} & & \\ & & \ddots & \\ & & & \lambda_{i,n_i}\end{pmatrix}\quad (i=1,\dots,p),$$

together with the conditions

$$\begin{aligned} &\lambda_{i,h},\ \lambda_{i,h}-\lambda_{i,h'}\notin \mathbf{Z}\quad (i=1,\dots,p,\ \ h,h'=1,\dots,n_i,\ \ h\neq h'),\\ &\rho_1-\rho_2\notin\mathbf{Z},\\ &\rho_j\notin\mathbf{Z}_{<0}\quad (j=1,2). \end{aligned}\tag{1.1}$$

1.1. Fundamental set of solutions. We have shown the following.

PROPOSITION 1. *The system* (0.4) *is completely integrable. Namely, we have* $d\,\Omega=\Omega\wedge\Omega$, *where* Ω *denotes the coefficient* 1*-form of* (0.4)*.*

This proposition follows from the relation (0.3) and guarantees the existence of n linearly independent solutions of the system (0.4).

PROPOSITION 2. *For each* i $(i=1,\dots,p)$, *the system* (0.4) *has* n_i *solutions*

$$Z_{i,h}(x,y)=(x-t_i)^{\lambda_{i,h}}(y-t_i)^{-\rho_1-\rho_2}W_{i,h}(x,y)\quad (h=1,\dots,n_i)$$

such that

$$W_{i,h}(t_i,y)=\textit{the } (n_1+\cdots+n_{i-1}+h)\textit{th unit } n\textit{-vector}\quad (h=1,\dots,n_i)$$

and $W_{i,h}(x,y)$ $(h=1,\dots,n_i)$ *are holomorphic in* $\mathcal{D}_i\times\Delta_i$, *where* $\mathcal{D}_i$ *is an arbitrary simply connected domain in* $\mathbf{C}$ *containing* t_i *and not containing* t_k $(k\neq i)$, *and* $\Delta_i=\mathbf{C}\setminus\overline{\mathcal{D}_i}$.

We set

$$\mathcal{Z}(x,y)=\Big(Z_{1,1}(x,y)\ \cdots\ Z_{1,n_1}(x,y)\ \cdots\ Z_{p,1}(x,y)\ \cdots\ Z_{p,n_p}(x,y)\Big).$$

PROPOSITION 3.

$$\begin{aligned}&\det\mathcal{Z}(x,y)\\ &=\frac{\prod_{i=1}^{p}\prod_{h=1}^{n_i}\Gamma(\lambda_{i,h}+1)}{\prod_{j=1}^{2}\Gamma(\rho_j+1)^{m_j}}\prod_{i=1}^{p}\prod_{h=1}^{n_i}\left\{(x-t_i)^{\lambda_{i,h}}(y-t_i)^{\lambda_{i,h}-\rho_1-\rho_2}(y-x)^{-\lambda_{i,h}}\right\}\end{aligned}$$

holds. Therefore $\mathcal{Z}(x,y)$ *is a fundamental set of solutions of* (0.4)*.*

1.2. Circuit matrices. We set

$$S_p=\bigcup_{i=1}^{p}\big(\{(x,y)\mid x=t_i\}\bigcup\{(x,y)\mid y=t_i\}\big)\bigcup\{(x,y)\mid x=y\}$$

and fix a base point (x_0,y_0) in $\mathbf{C}^2\setminus S_p$. Let $\mathcal{C}_x$ be a simple closed curve on the (x,y_0)-plane ($=\{(x,y_0)\mid x\in\mathbf{C}\}$) such that the points (x_0,y_0), (y_0,y_0), (t_1,y_0), $(t_2,y_0),\dots,(t_p,y_0)$ lie on $\mathcal{C}_x$ and come in this order when we trace $\mathcal{C}_x$ in the positive direction. For each i $(i=0,1,\dots,p)$, let σ_i be a loop on the (x,y_0)-plane that starts from (x_0,y_0), goes inside $\mathcal{C}_x$, encircles (t_i,y_0) once in the positive direction, and returns inside $\mathcal{C}_x$ to (x_0,y_0), where we set $t_0=y_0$. Let $\mathcal{C}_y$ be a copy of $\mathcal{C}_x$ onto the (x_0,y)-plane ($=\{(x_0,y)\mid y\in\mathbf{C}\}$); that is, $\mathcal{C}_y$ is a curve on the (x_0,y)-plane such that the points (x_0,x_0), (x_0,y_0), (x_0,t_1), $(x_0,t_2),\dots,(x_0,t_p)$ lie on $\mathcal{C}_y$ and come in this order when we trace $\mathcal{C}_y$ in the positive direction. Let $\mathcal{L}$ be a simple curve on the (x_0,y)-plane that starts from a point located between (x_0,y_0) and (x_0,t_1) on $\mathcal{C}_y$, goes outside $\mathcal{C}_y$, and ends at (x_0,∞). For each j $(j=1,\dots,p,p+1)$, let τ_j be a loop on the (x_0,y)-plane that starts from (x_0,y_0), goes outside $\mathcal{C}_y$ not crossing $\mathcal{L}$, encircles

(x_0, t_j) once in the positive direction, and returns outside $\mathcal{C}_y$ to (x_0, y_0) not crossing $\mathcal{L}$, where we set $t_{p+1} = x_0$. Then the homotopy classes $[\sigma_i]$ $(i = 0, 1, \dots, p)$ and $[\tau_j]$ $(j = 1, \dots, p, p+1)$ generate the fundamental group $\pi_1(\mathbf{C}^2 \setminus S_p, (x_0, y_0))$.

If we continue analytically the fundamental set of solutions $\mathcal{Z}(x, y)$ along the loop σ_i (resp., τ_j), then we obtain another fundamental set of solutions $\mathcal{Z}(x, y)M_i$ (resp., $\mathcal{Z}(x, y)N_j$), where M_i (resp., N_j) is an element in $GL(n, \mathbf{C})$. We call M_i (resp., N_j) the *circuit matrix* of $\mathcal{Z}(x, y)$ along σ_i (resp., τ_j). The *monodromy group* of the system (0.4) with respect to $\mathcal{Z}(x, y)$ is a subgroup of $GL(n, \mathbf{C})$ generated by M_i $(i = 0, 1, \dots, p)$ and N_j $(j = 1, \dots, p, p+1)$, that is to say, an image of the representation

$$R : \pi_1(\mathbf{C}^2 \setminus S_p, (x_0, y_0)) \longrightarrow GL(n, \mathbf{C})$$

that is defined by

$$\mathcal{Z}(x, y)^\gamma = \mathcal{Z}(x, y) \cdot R([\gamma])$$

for any loop γ in $\mathbf{C}^2 \setminus S_p$ with the base point (x_0, y_0), where $\mathcal{Z}(x, y)^\gamma$ denotes the analytic continuation of $\mathcal{Z}(x, y)$ along the loop γ.

We have shown the following.

Proposition 4.

(i) *For $i = 1, \dots, p$, the circuit matrix M_i has the form*

$$M_i = \begin{pmatrix} I_{n_1} & & & & & & \\ & \ddots & & & & & \\ & & I_{n_{i-1}} & & & & \\ M_{i1} & \cdots & M_{i,i-1} & M_{ii} & M_{i,i+1} & \cdots & M_{ip} \\ & & & & I_{n_{i+1}} & & \\ & & & & & \ddots & \\ & & & & & & I_{n_p} \end{pmatrix},$$

where

$$M_{ii} = \begin{pmatrix} e(\lambda_{i,1}) & & & \\ & e(\lambda_{i,2}) & & \\ & & \ddots & \\ & & & e(\lambda_{i,n_i}) \end{pmatrix}.$$

(ii) *For $j = 1, \dots, p$, the circuit matrix N_j has the form*

$$N_j = \begin{pmatrix} I_{n_1} & & & N_{1j} & & & \\ & \ddots & & \vdots & & & \\ & & I_{n_{j-1}} & N_{j-1,j} & & & \\ & & & N_{jj} & & & \\ & & & N_{j+1,j} & I_{n_{j+1}} & & \\ & & & \vdots & & \ddots & \\ & & & N_{pj} & & & I_{n_p} \end{pmatrix}.$$

(iii) *The Jordan canonical form of the circuit matrix M_0 is*

$$\begin{pmatrix} e(-\rho_1)I_{m_1} & \\ & e(-\rho_2)I_{m_2} \end{pmatrix}.$$

Proposition 5. *The circuit matrices M_i $(i = 0, 1, \dots, p)$ and N_j $(j = 1, \dots, p, p + 1)$ have the following relations*:

$$M_p M_{p-1} \cdots M_1 M_0 = I_n, \tag{1.2}$$

$$N_1 N_2 \cdots N_p N_{p+1} = e(-\rho_1 - \rho_2) \cdot I_n, \tag{1.3}$$

$$N_{p+1} = M_0, \tag{1.4}$$

and

$$M_i N_j = N_j M_i \tag{1.5}$$

for $i, j = 1, \dots, p$ with $i \neq j$.

Remark. From the relations (1.2) and (1.4) it follows that the monodromy group of (0.4) is generated by M_i $(i = 1, \dots, p)$ and N_j $(j = 1, \dots, p)$.

Restricting the system (0.4) to the (x, y_0)-plane, we obtain

$$\frac{dZ}{dx} = \Big((xI_n - T)^{-1} - \frac{1}{x - y_0} I_n\Big) AZ,$$

which is transformed into the system of Okubo normal form

$$(tI_n - T')\frac{dZ}{dt} = AZ, \tag{1.6}$$

where $T' = (T - y_0 I_n)^{-1}$, by the change of the variable x for $t = (x - y_0)^{-1}$. Since the matrices M_i $(i = 1, \dots, p)$ are also circuit matrices for the system (1.6), we can determine M_i $(i = 1, \dots, p)$ explicitly up to a diagonal transformation if (1.6) is free from accessory parameters. Once the M_i's are determined, we can determine the matrices N_j $(j = 1, \dots, p)$ by the relation

$$N_{ij} = [e(-\rho_1 - \rho_2) \cdot M_{j-1} M_{j-2} \cdots M_1 M_p M_{p-1} \cdots M_j]_{ij} \tag{1.7}$$

for $i, j = 1, \dots, p$, where $[\quad]_{ij}$ denotes the (i, j)-block in the $(n_1, n_2, \dots, n_p)$-decomposition of an $n \times n$ matrix. The relation (1.7) is derived from the form of N_j and the relations (1.2)–(1.5) (see Lemma 1 in section 2.2).

2. System (I**).

Let n be an integer equal to or greater than 3, and let t_i $(i = 1, \dots, n)$ be mutually distinct points in $\mathbf{C}$. We are concerned with a representation of the fundamental group of $\mathbf{C}^2 \setminus S_n$, where

$$S_n = \bigcup_{i=1}^{n} \big(\{(x, y) \mid x = t_i\} \bigcup \{(x, y) \mid y = t_i\}\big) \bigcup \{(x, y) \mid x = y\}.$$

2.1. Monodromy group of system (I**).

Let $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ be an element in $\mathbf{C}^n$, and let ρ_1 be an element in $\mathbf{C}$. We set

$$\rho_2 = \sum_{i=1}^{n} \lambda_i - (n - 1)\rho_1 \quad \text{and} \quad \rho = (\rho_1, \rho_2).$$

System $(\mathrm{I}^{**})_{\lambda,\rho}$ (or simply (I^{**})) of rank n is the system of total differential equations

(2.1)

$$dZ = \Big\{(xI_n - T_{\mathrm{I}^*})^{-1} A_{\mathrm{I}^*} dx + (A_{\mathrm{I}^*} - (\rho_1 + \rho_2) I_n)(yI_n - T_{\mathrm{I}^*})^{-1} dy - A_{\mathrm{I}^*} \frac{d(x - y)}{x - y}\Big\} Z$$

for an unknown n-dimensional column vector Z, where

$$T_{\mathrm{I}^*} = \begin{pmatrix} t_1 & & & \\ & t_2 & & \\ & & \ddots & \\ & & & t_n \end{pmatrix} \quad \text{and} \quad A_{\mathrm{I}^*} = \begin{pmatrix} \lambda_1 & \lambda_1-\rho_1 & \cdots & \lambda_1-\rho_1 \\ \lambda_2-\rho_1 & \lambda_2 & \cdots & \lambda_2-\rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n-\rho_1 & \lambda_n-\rho_1 & \cdots & \lambda_n \end{pmatrix}.$$

Under the condition that $\rho_1 \neq \rho_2$, the Jordan canonical form of A_{I^*} is

$$\begin{pmatrix} \rho_1 I_{n-1} & \\ & \rho_2 \end{pmatrix}$$

(see Haraoka [2, Thm. I*]). Therefore, A_{I^*} satisfies the relation

$$(A_{\mathrm{I}^*} - \rho_1 I_n)(A_{\mathrm{I}^*} - \rho_2 I_n) = O,$$

and hence the system (2.1) is completely integrable.

We assume the conditions

$$\begin{aligned} \lambda_i &\notin \mathbf{Z} \quad (i = 1, \dots, n), \\ \rho_1 - \rho_2 &\notin \mathbf{Z}, \\ \rho_j &\notin \mathbf{Z}_{<0} \quad (j = 1, 2), \end{aligned} \tag{2.2}$$

which are corresponding to (1.1). Then the system (2.1) has a fundamental set of solutions such as we state in section 1.1, which we denote by $\mathcal{Z}_{\mathrm{I}^{**}}(x, y)$.

Theorem 1. *We assume* (2.2) *and*

$$\lambda_i - \rho_1 \notin \mathbf{Z} \quad (i = 1, \dots, n).$$

There is a diagonal matrix $D \in GL(n, \mathbf{C})$ such that the monodromy group of the system (2.1) *with respect to $\mathcal{Z}_{\mathrm{I}^{**}}(x, y)D$ is generated by*

$$M_i = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ \xi_{i1} & \cdots & \xi_{i,i-1} & e(\lambda_i) & \xi_{i,i+1} & \cdots & \xi_{in} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \quad (i = 1, \dots, n)$$

and

$$N_j = \begin{pmatrix} 1 & & & \eta_{1j} & & & \\ & \ddots & & \vdots & & & \\ & & 1 & \eta_{j-1,j} & & & \\ & & & \eta_{jj} & & & \\ & & & \eta_{j+1,j} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & \eta_{nj} & & & 1 \end{pmatrix} \quad (j = 1, \dots, n)$$

with

$$\xi_{ij} = \begin{cases} e(\lambda_j) - e(\rho_1) & \text{for } i,j = 1, \dots, n \text{ with } i < j, \\ e(\lambda_j - \rho_1) - 1 & \text{for } i,j = 1, \dots, n \text{ with } i > j, \end{cases} \tag{2.3}$$

and

(2.4)

$$\begin{aligned} \eta_{jj} &= e(\lambda_j - \rho_1 - \rho_2) && \text{for } j = 1, \dots, n, \\ \eta_{ij} &= e(\lambda_j + \lambda_{j+1} + \cdots + \lambda_{j+r-1} - r\rho_1 - \rho_2) \cdot \xi_{ij} && \text{for } i,j = 1, \dots, n \text{ with } i \neq j, \end{aligned}$$

where we have set

$$r = r(i,j) = \begin{cases} n + i - j & \text{for } i < j, \\ i - j & \text{for } i > j, \end{cases}$$

and

$$\lambda_\ell = \lambda_{\ell - n} \quad \text{for } \ell > n.$$

2.2. Proof of Theorem 1. We have only to prove (2.4), since the elements (2.3) of M_i $(i = 1, \dots, n)$ have already been evaluated by Takano and Bannai [8] and Haraoka [1]. Note that M_i $(i = 1, \dots, n)$ and N_j $(j = 1, \dots, n)$ satisfy

$$N_1 N_2 \cdots N_n = e(-\rho_1 - \rho_2) \cdot M_n M_{n-1} \cdots M_1 \tag{2.5}$$

and

$$N_j M_i = M_i N_j \tag{2.6}$$

for $i, j = 1, \dots, n$ with $i \neq j$.

Lemma 1. *For $i, j = 1, \dots, n$, we have*

$$\eta_{ij} = [e(-\rho_1 - \rho_2) \cdot M_{j-1} M_{j-2} \cdots M_1 M_n M_{n-1} \cdots M_j]_{ij},$$

where $[\quad]_{ij}$ denotes the (i,j)-element.

Proof. Combining (2.5) and (2.6) we obtain

$$N_j N_{j+1} \cdots N_n N_1 N_2 \cdots N_{j-1} = e(-\rho_1 - \rho_2) \cdot M_{j-1} M_{j-2} \cdots M_1 M_n M_{n-1} \cdots M_j.$$

By virtue of the form of the N_j's we obtain

$$[N_j N_{j+1} \cdots N_n N_1 N_2 \cdots N_{j-1}]_{ij} = \eta_{ij}.$$

Hence, we have the relation. This completes the proof. □

Proof of Theorem 1.

Step 1. We first prove (2.4) for $j = 1$ by induction on n. In the case $n = 3$ we can obtain

$$M_3M_2M_1 = \begin{pmatrix} e(\lambda_1) & * & * \\ e(\lambda_1)(e(\lambda_1-\rho_1)-1) & * & * \\ e(\lambda_1+\lambda_2-\rho_1)(e(\lambda_1-\rho_1)-1) & * & * \end{pmatrix}$$

by direct calculation. Applying Lemma 1 with $j=1$, we have

$$\begin{aligned} \eta_{11} &= e(\lambda_1-\rho_1-\rho_2), \\ \eta_{i1} &= e(\lambda_1+\lambda_{i-1}-(i-1)\rho_1-\rho_2)\cdot\xi_{i1} \quad (i=2,3), \end{aligned}$$

which shows that (2.4) is valid for $j=1$ in the case $n=3$.

We assume for induction that (2.4) is valid for $j=1$ in the case of rank $n-1$. Then, we shall prove (2.4) for $j=1$ in the case of rank n. We use the notation λ'_i, ρ'_j, M'_i, and η'_{ij} for the case of rank $n-1$ instead of λ_i, ρ_j, M_i, and η_{ij} for the case of rank n, respectively. We set

$$L = \bigl(\zeta_{ij}\bigr)_{\substack{1\le i\le n\\1\le j\le n}} = M_nM_{n-1}\cdots M_1 \quad\text{and}\quad L' = \bigl(\zeta'_{ij}\bigr)_{\substack{1\le i\le n-1\\1\le j\le n-1}} = M'_{n-1}M'_{n-2}\cdots M'_1.$$

Provided that $\lambda'_i=\lambda_i$ for $i=1,\dots,n-1$ and $\rho'_1=\rho_1$, we have

$$M_i = \left(\begin{array}{ccc|c} & M'_i & & * \\ \hline 0 & \cdots & 0 & 1 \end{array}\right)$$

for $i=1,\dots,n-1$, and hence

$$L = M_n\left(\begin{array}{ccc|c} & L' & & * \\ \hline 0 & \cdots & 0 & 1 \end{array}\right) = \left(\begin{array}{ccc|c} & & & 0 \\ & I_{n-1} & & \vdots \\ & & & 0 \\ \hline \xi_{n1} & \cdots & \xi_{n,n-1} & e(\lambda_n) \end{array}\right)\left(\begin{array}{ccc|c} & L' & & * \\ \hline 0 & \cdots & 0 & 1 \end{array}\right)$$

From this expression it follows that

$$\zeta_{i1} = \zeta'_{i1} \quad (i=1,\dots,n-1) \quad\text{and}\quad \zeta_{n1} = \sum_{k=1}^{n-1}\xi_{nk}\cdot\zeta'_{k1}.$$

Using these relations and the assumption for induction with Lemma 1, we therefore obtain

$$
\begin{aligned}
\eta_{11} &= e(-\rho_1-\rho_2)\cdot\zeta_{11} = e(-\rho_1-\rho_2)\cdot\zeta'_{11} = e(-\rho_1-\rho_2)\cdot e(\rho_1+\rho'_2)\cdot\eta'_{11}\\
&= e(\rho'_2-\rho_2)\cdot e(\lambda_1-\rho_1-\rho'_2)\\
&= e(\lambda_1-\rho_1-\rho_2),\\
\eta_{i1} &= e(-\rho_1-\rho_2)\cdot\zeta_{i1} = e(-\rho_1-\rho_2)\cdot\zeta'_{i1} = e(-\rho_1-\rho_2)\cdot e(\rho_1+\rho'_2)\cdot\eta'_{i1}\\
&= e(\rho'_2-\rho_2)\cdot e(\lambda_1+\lambda_2+\cdots+\lambda_{i-1}-(i-1)\rho_1-\rho'_2)\cdot\xi_{i1}\\
&= e(\lambda_1+\lambda_2+\cdots+\lambda_{i-1}-(i-1)\rho_1-\rho_2)\cdot\xi_{i1}\quad(i=2,\dots,n-1),\\
\eta_{n1} &= e(-\rho_1-\rho_2)\cdot\zeta_{n1} = e(-\rho_1-\rho_2)\cdot\sum_{k=1}^{n-1}\xi_{nk}\cdot e(\rho_1+\rho'_2)\cdot\eta'_{k1}\\
&= e(\rho'_2-\rho_2)\cdot\Big\{(e(\lambda_1-\rho_1)-1)\cdot e(\lambda_1-\rho_1-\rho'_2)\\
&\quad+\sum_{k=2}^{n-1}(e(\lambda_k-\rho_1)-1)\cdot e(\lambda_1+\cdots+\lambda_{k-1}-(k-1)\rho_1-\rho'_2)\cdot(e(\lambda_1-\rho_1)-1)\Big\}\\
&= e(\rho'_2-\rho_2)\cdot(e(\lambda_1-\rho_1)-1)\cdot\Big\{e(\lambda_1-\rho_1-\rho'_2)\\
&\quad+\sum_{k=2}^{n-1}\Big(e(\lambda_1+\cdots+\lambda_k-k\rho_1-\rho'_2)-e(\lambda_1+\cdots+\lambda_{k-1}-(k-1)\rho_1-\rho'_2)\Big)\Big\}\\
&= e(\rho'_2-\rho_2)\cdot(e(\lambda_1-\rho_1)-1)\cdot e(\lambda_1+\lambda_2+\cdots+\lambda_{n-1}-(n-1)\rho_1-\rho'_2)\\
&= e(\lambda_1+\lambda_2+\cdots+\lambda_{n-1}-(n-1)\rho_1-\rho_2)\cdot\xi_{n1},
\end{aligned}
$$

which shows that (2.4) is valid for $j=1$ in the case of rank n.

Step 2. We next prove (2.4) for $j=2$. We set

$$
R=\begin{pmatrix}
0 & & & e(\rho_1)\\
1 & \ddots & & \\
 & \ddots & \ddots & \\
 & & 1 & 0
\end{pmatrix}.
$$

Moreover, we set

$$
\tilde{M}_i = R^{-1}M_{i+1}R = \begin{pmatrix}
1 & & & & & \\
 & \ddots & & & & \\
\xi_{i+1,2} & \cdots & e(\lambda_{i+1}) & \cdots & \xi_{i+1,n} & e(\rho_1)\xi_{i+1,1}\\
 & & & \ddots & & \\
 & & & & \ddots & \\
 & & & & & 1
\end{pmatrix}
$$

for $i=1,2,\dots,n-1$, and

$$
\tilde{M}_n = R^{-1}M_1R = \begin{pmatrix}
1 & & & \\
 & \ddots & & \\
 & & 1 & \\
e(-\rho_1)\xi_{12} & \cdots & e(-\rho_1)\xi_{1,n-1} & e(\lambda_1)
\end{pmatrix}.
$$

Here we see that for each i ($i = 1, \dots, n$), $\tilde{M}_i$ agrees with M_i with a parameter set $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ replaced by $\tilde{\lambda} = (\lambda_2, \lambda_3, \dots, \lambda_n, \lambda_1)$. Hence, for each i ($i = 1, \dots, n$), the $(i, 1)$-element of the matrix $e(-\rho_1 - \rho_2) \cdot \tilde{M}_n \tilde{M}_{n-1} \cdots \tilde{M}_1$ is equal to η_{i1} with $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ replaced by $\tilde{\lambda} = (\lambda_2, \lambda_3, \dots, \lambda_n, \lambda_1)$, that is,

$$\begin{aligned}
&\big[e(-\rho_1 - \rho_2) \cdot \tilde{M}_n \tilde{M}_{n-1} \cdots \tilde{M}_1\big]_{i1} \\
&\quad = \begin{cases} e(\lambda_2 - \rho_1 - \rho_2) & (i = 1), \\ e(\lambda_2 + \lambda_3 + \cdots + \lambda_i - (i-1)\rho_1 - \rho_2) \cdot (e(\lambda_2 - \rho_1) - 1) & (i = 2, \dots, n). \end{cases}
\end{aligned}$$

On the other hand, by Lemma 1 with $j = 2$, we have

$$\begin{aligned}
\eta_{i2} &= \big[e(-\rho_1 - \rho_2) \cdot M_1 M_n M_{n-1} \cdots M_2\big]_{i2} \\
&= \big[e(-\rho_1 - \rho_2) \cdot R \tilde{M}_n \tilde{M}_{n-1} \cdots \tilde{M}_1 R^{-1}\big]_{i2} \\
&= \begin{cases} e(\rho_1) \cdot \big[e(-\rho_1 - \rho_2) \cdot \tilde{M}_n \tilde{M}_{n-1} \cdots \tilde{M}_1\big]_{n1} & (i = 1), \\ \big[e(-\rho_1 - \rho_2) \cdot \tilde{M}_n \tilde{M}_{n-1} \cdots \tilde{M}_1\big]_{i-1,1} & (i = 2, \dots, n). \end{cases}
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\eta_{12} &= e(\rho_1) \cdot e(\lambda_2 + \cdots + \lambda_n - (n-1)\rho_1 - \rho_2) \cdot (e(\lambda_2 - \rho_1) - 1) \\
&= e(\lambda_2 + \cdots + \lambda_n - (n-1)\rho_1 - \rho_2) \cdot \xi_{12}, \\
\eta_{22} &= e(\lambda_2 - \rho_1 - \rho_2), \\
\eta_{i2} &= e(\lambda_2 + \cdots + \lambda_{i-1} - (i-2)\rho_1 - \rho_2) \cdot (e(\lambda_2 - \rho_1) - 1) \\
&= e(\lambda_2 + \cdots + \lambda_{i-1} - (i-2)\rho_1 - \rho_2) \cdot \xi_{i2} \quad (i = 3, \dots, n),
\end{aligned}$$

which shows that (2.4) is valid for $j = 2$.

Step 3. For $j = 3, 4, \dots, n$, we can prove (2.4) by the same consideration as developed in Step 2 with the matrices $R^{-(j-1)} M_i R^{j-1}$ ($i = 1, \dots, n$). Thus the proof of Theorem 1 is complete. □

3. System (II**), system (III**), and system (IV**).

Let t_1, t_2, and t_3 be mutually distinct points in $\mathbf{C}$. We are concerned with representations of the fundamental group of $\mathbf{C}^2 \setminus S_3$, where

$$S_3 = \bigcup_{i=1}^{3} \big(\{(x, y) \mid x = t_i\} \bigcup \{(x, y) \mid y = t_i\}\big) \bigcup \{(x, y) \mid x = y\}.$$

3.1. Monodromy group of system (II**).

Let n be an even integer equal to or greater than 4. We set $n = 2m$ with $m \in \{2, 3, 4, \dots\}$. Let $\lambda = (\lambda_1, \dots, \lambda_m)$, $\mu = (\mu_1, \dots, \mu_{m-1})$, ν, and $\rho = (\rho_1, \rho_2)$ be elements in $\mathbf{C}^m$, $\mathbf{C}^{m-1}$, $\mathbf{C}$, and $\mathbf{C}^2$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m-1} \mu_i + \nu = m\rho_1 + m\rho_2.$$

System $(\mathrm{II}^{**})_{\lambda,\mu,\nu,\rho}$ (or simply (II^{**})) of rank n is the system of total differential equations

(3.1)

$$dZ = \Big\{(xI_n - T_{\mathrm{II}^*})^{-1}A_{\mathrm{II}^*}dx + (A_{\mathrm{II}^*} - (\rho_1+\rho_2)I_n)(yI_n - T_{\mathrm{II}^*})^{-1}dy - A_{\mathrm{II}^*}\frac{d(x-y)}{x-y}\Big\}Z$$

with

$$T_{\mathrm{II}^*} = \begin{pmatrix} t_1 I_m & & \\ & t_2 I_{m-1} & \\ & & t_3 \end{pmatrix},$$

$$A_{\mathrm{II}^*} = \left(\begin{array}{ccc|cccc}
\lambda_1 & & & & & & \\
& \ddots & & & & (\alpha_{ij}) & \\
& & \lambda_m & & & & \\ \hline
& & & \mu_1 & & & \gamma_1 \\
& (\beta_{ij}) & & & \ddots & & \vdots \\
& & & & & \mu_{m-1} & \gamma_{m-1} \\
& & & \delta_1 & \cdots & \delta_{m-1} & \nu
\end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1)(\lambda_i - \rho_2)\cdot \prod_{k\in\{1,\dots,m\}\setminus\{i\}} \frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_k}$$
$$(i = 1,\dots,m,\ \ j = 1,\dots,m-1),$$
$$\alpha_{im} = (\lambda_i - \rho_1)(\lambda_i - \rho_2)\cdot \prod_{k\in\{1,\dots,m\}\setminus\{i\}} \frac{1}{\lambda_i - \lambda_k} \quad (i = 1,\dots,m),$$
$$\beta_{ij} = \prod_{\ell\in\{1,\dots,m-1\}\setminus\{i\}} \frac{\lambda_j + \mu_\ell - \rho_1 - \rho_2}{\mu_i - \mu_\ell} \quad (i = 1,\dots,m-1,\ \ j = 1,\dots,m),$$
$$\beta_{mj} = -\prod_{\ell\in\{1,\dots,m-1\}} (\lambda_j + \mu_\ell - \rho_1 - \rho_2) \quad (j = 1,\dots,m),$$
$$\gamma_i = \prod_{\ell\in\{1,\dots,m-1\}\setminus\{i\}} \frac{1}{\mu_i - \mu_\ell} \quad (i = 1,\dots,m-1),$$
$$\delta_j = -\prod_{k\in\{1,\dots,m\}} (\lambda_k + \mu_j - \rho_1 - \rho_2) \quad (j = 1,\dots,m-1).$$

The Jordan canonical form of A_{II^*} is

$$\begin{pmatrix} \rho_1 I_m & \\ & \rho_2 I_m \end{pmatrix}$$

(see Haraoka [2, Thm. II*]). Therefore A_{II^*} satisfies the relation

$$(A_{\mathrm{II}^*} - \rho_1 I_n)(A_{\mathrm{II}^*} - \rho_2 I_n) = O,$$

and hence the system (3.1) is completely integrable.

We assume the conditions

$$
\begin{aligned}
\lambda_i,\ \lambda_i-\lambda_j &\notin \mathbf{Z} \quad (i,j=1,\dots,m,\ \ i\neq j),\\
\mu_i,\ \mu_i-\mu_j &\notin \mathbf{Z} \quad (i,j=1,\dots,m-1,\ \ i\neq j),\\
\nu &\notin \mathbf{Z},\\
\rho_1-\rho_2 &\notin \mathbf{Z},\\
\rho_j &\notin \mathbf{Z}_{<0} \quad (j=1,2),
\end{aligned}
\tag{3.2}
$$

which are corresponding to (1.1). Then the system (3.1) has a fundamental set of solutions such as we state in section 1.1, which we denote by $\mathcal{Z}_{\mathrm{II}^{**}}(x,y)$.

Theorem 2. *We assume* (3.2) *and*

$$
\begin{aligned}
\lambda_i-\rho_k &\notin \mathbf{Z} \quad (i=1,\dots,m,\ \ k=1,2),\\
\lambda_i+\mu_j-\rho_1-\rho_2 &\notin \mathbf{Z} \quad (i=1,\dots,m,\ \ j=1,\dots,m-1).
\end{aligned}
$$

There is a diagonal matrix $D\in GL(n,\mathbf{C})$ such that the monodromy group of the system (3.1) *with respect to $\mathcal{Z}_{\mathrm{II}^{**}}(x,y)D$ is generated by*

$$
\begin{aligned}
M_1 &= \begin{pmatrix} E_m(\lambda) & (\xi_{ij})_{\substack{1\le i\le m\\ 1\le j\le m}} \\ O & I_m \end{pmatrix},\\
M_2 &= \begin{pmatrix} I_m & O & O \\ (\eta_{ij})_{\substack{1\le i\le m-1\\ 1\le j\le m}} & E_{m-1}(\mu) & (\eta_{in})_{1\le i\le m-1} \\ O & O & 1 \end{pmatrix},\\
M_3 &= \begin{pmatrix} I_{n-1} & O \\ (\zeta_j)_{1\le j\le n-1} & e(\nu) \end{pmatrix},\\
N_1 &= \begin{pmatrix} E_m(\lambda-\rho_1-\rho_2) & O \\ (\theta_{ij})_{\substack{1\le i\le m\\ 1\le j\le m}} & I_m \end{pmatrix},\\
N_2 &= \begin{pmatrix} I_m & (\phi_{ij})_{\substack{1\le i\le m\\ 1\le j\le m-1}} & O \\ O & E_{m-1}(\mu-\rho_1-\rho_2) & O \\ O & (\phi_{nj})_{1\le j\le m-1} & 1 \end{pmatrix},\\
N_3 &= \begin{pmatrix} I_{n-1} & (\psi_i)_{1\le i\le n-1} \\ O & e(\nu-\rho_1-\rho_2) \end{pmatrix},
\end{aligned}
$$

where

$$
E_m(\lambda)=\begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_m) \end{pmatrix}, \qquad E_m(\lambda-\rho_1-\rho_2)=e(-\rho_1-\rho_2)E_m(\lambda),
$$

$$
E_{m-1}(\mu)=\begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_{m-1}) \end{pmatrix}, \qquad E_{m-1}(\mu-\rho_1-\rho_2)=e(-\rho_1-\rho_2)E_{m-1}(\mu),
$$

$$\xi_{ij} = (e(\lambda_i) - e(\rho_1))(e(\rho_2 - \lambda_i) - 1) \cdot \prod_{k \in \{1,\dots,m\} \setminus \{i\}} \frac{e(\mu_j) - e(\rho_1 + \rho_2 - \lambda_k)}{e(\rho_1 + \rho_2 - \lambda_i) - e(\rho_1 + \rho_2 - \lambda_k)}$$
$$(i = 1, \dots, m, \;\; j = 1, \dots, m-1),$$

$$\xi_{im} = (e(\lambda_i) - e(\rho_1))(e(\rho_2 - \lambda_i) - 1) \cdot \prod_{k \in \{1,\dots,m\} \setminus \{i\}} \frac{1}{e(\rho_1 + \rho_2 - \lambda_i) - e(\rho_1 + \rho_2 - \lambda_k)}$$
$$(i = 1, \dots, m),$$

$$\eta_{ij} = \prod_{\ell \in \{1,\dots,m-1\} \setminus \{i\}} \frac{e(\rho_1 + \rho_2 - \lambda_j) - e(\mu_\ell)}{e(\mu_i) - e(\mu_\ell)} \qquad (i = 1, \dots, m-1, \;\; j = 1, \dots, m),$$

$$\eta_{in} = \prod_{\ell \in \{1,\dots,m-1\} \setminus \{i\}} \frac{1}{e(\mu_i) - e(\mu_\ell)} \qquad (i = 1, \dots, m-1),$$

$$\zeta_j = e(\lambda_j + \nu - \rho_1 - \rho_2) \cdot \prod_{\ell \in \{1,\dots,m-1\}} (e(\rho_1 + \rho_2 - \lambda_j) - e(\mu_\ell)) \qquad (j = 1, \dots, m),$$

$$\zeta_{m+j} = -\frac{1}{e(\mu_j)} \cdot \prod_{k \in \{1,\dots,m\}} (e(\mu_j) - e(\rho_1 + \rho_2 - \lambda_k)) \qquad (j = 1, \dots, m-1),$$

and

$$\begin{aligned}
\theta_{ij} &= e(\lambda_j - \rho_1 - \rho_2) \cdot \eta_{ij} \qquad (i = 1, \dots, m-1, \;\; j = 1, \dots, m), \\
\theta_{mj} &= e(-\nu) \cdot \zeta_j \qquad (j = 1, \dots, m), \\
\phi_{ij} &= e(-\lambda_i) \cdot \xi_{ij} \qquad (i = 1, \dots, m, \;\; j = 1, \dots, m-1), \\
\phi_{nj} &= e(\mu_j - \rho_1 - \rho_2) \cdot \zeta_{m+j} \qquad (j = 1, \dots, m-1), \\
\psi_i &= e(\nu - \rho_1 - \rho_2) \cdot \xi_{im} \qquad (i = 1, \dots, m), \\
\psi_{m+i} &= e(-\mu_i) \cdot \eta_{in} \qquad (i = 1, \dots, m-1).
\end{aligned}$$

3.2. Monodromy group of system (III).** Let n be an odd integer equal to or greater than 5. We set $n = 2m+1$ with $m \in \{2, 3, 4, \dots\}$. Let $\lambda = (\lambda_1, \dots, \lambda_m)$, $\mu = (\mu_1, \dots, \mu_m)$, ν, and $\rho = (\rho_1, \rho_2)$ be elements in $\mathbf{C}^m$, $\mathbf{C}^m$, $\mathbf{C}$, and $\mathbf{C}^2$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i + \nu = (m+1)\rho_1 + m\rho_2.$$

System $(\mathrm{III}^{**})_{\lambda,\mu,\nu,\rho}$ (or simply (III**)) of rank n is the system of total differential equations

(3.3)

$$dZ = \Big\{ (xI_n - T_{\mathrm{III}^*})^{-1} A_{\mathrm{III}^*} dx + (A_{\mathrm{III}^*} - (\rho_1 + \rho_2) I_n)(yI_n - T_{\mathrm{III}^*})^{-1} dy - A_{\mathrm{III}^*} \frac{d(x-y)}{x-y} \Big\} Z$$

with

$$T_{\mathrm{III}^*} = \begin{pmatrix} t_1 I_m & & \\ & t_2 & \\ & & t_3 I_m \end{pmatrix},$$

$$A_{\mathrm{III}^*} = \begin{pmatrix} \lambda_1 & & & \gamma_1 & & & \\ & \ddots & & \vdots & & (\alpha_{ij}) & \\ & & \lambda_m & \gamma_m & & & \\ \epsilon_1 & \cdots & \epsilon_m & \nu & \kappa_1 & \cdots & \kappa_m \\ & & & \delta_1 & \mu_1 & & \\ & (\beta_{ij}) & & \vdots & & \ddots & \\ & & & \delta_m & & & \mu_m \end{pmatrix},$$

where

$$\begin{aligned} \alpha_{ij} &= (\lambda_i - \rho_1) \cdot \prod_{k \in \{1,\dots,m\} \setminus \{i\}} \frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_k - \lambda_i} \quad (i,j = 1,\dots,m), \\ \beta_{ij} &= (\mu_i - \rho_1) \cdot \prod_{\ell \in \{1,\dots,m\} \setminus \{i\}} \frac{\rho_1 + \rho_2 - \lambda_j - \mu_\ell}{\mu_i - \mu_\ell} \quad (i,j = 1,\dots,m), \\ \gamma_i &= (\lambda_i - \rho_1) \cdot \prod_{k \in \{1,\dots,m\} \setminus \{i\}} \frac{1}{\lambda_k - \lambda_i} \quad (i = 1,\dots,m), \\ \delta_i &= (\mu_i - \rho_1) \cdot \prod_{\ell \in \{1,\dots,m\} \setminus \{i\}} \frac{1}{\mu_i - \mu_\ell} \quad (i = 1,\dots,m), \\ \epsilon_j &= \prod_{\ell \in \{1,\dots,m\}} (\rho_1 + \rho_2 - \lambda_j - \mu_\ell) \quad (j = 1,\dots,m), \\ \kappa_j &= -\prod_{k \in \{1,\dots,m\}} (\lambda_k + \mu_j - \rho_1 - \rho_2) \quad (j = 1,\dots,m). \end{aligned}$$

The Jordan canonical form of A_{III^*} is

$$\begin{pmatrix} \rho_1 I_{m+1} & \\ & \rho_2 I_m \end{pmatrix}$$

(see Haraoka [2, Thm. III*]). Therefore A_{III^*} satisfies the relation

$$(A_{\mathrm{III}^*} - \rho_1 I_n)(A_{\mathrm{III}^*} - \rho_2 I_n) = O,$$

and hence the system (3.3) is completely integrable.

We assume the conditions

$$\begin{aligned} \lambda_i,\ \lambda_i - \lambda_j &\notin \mathbf{Z} \quad (i,j = 1,\dots,m,\ i \neq j), \\ \mu_i,\ \mu_i - \mu_j &\notin \mathbf{Z} \quad (i,j = 1,\dots,m,\ i \neq j), \\ \nu &\notin \mathbf{Z}, \\ \rho_1 - \rho_2 &\notin \mathbf{Z}, \\ \rho_j &\notin \mathbf{Z}_{<0} \quad (j = 1,2), \end{aligned} \tag{3.4}$$

which are corresponding to (1.1). Then the system (3.3) has a fundamental set of solutions such as we state in section 1.1, which we denote by $\mathcal{Z}_{\mathrm{III}^{**}}(x,y)$.

Theorem 3. *We assume* (3.4) *and*

$$\begin{aligned} \lambda_i - \rho_1 &\notin \mathbf{Z} \quad (i = 1,\dots,m), \\ \mu_i - \rho_1 &\notin \mathbf{Z} \quad (i = 1,\dots,m), \\ \lambda_i + \mu_j - \rho_1 - \rho_2 &\notin \mathbf{Z} \quad (i,j = 1,\dots,m). \end{aligned}$$

There is a diagonal matrix $D \in GL(n, \mathbf{C})$ such that the monodromy group of the system (3.3) *with respect to $\mathcal{Z}_{\mathrm{III}^{**}}(x, y)D$ is generated by*

$$M_1 = \begin{pmatrix} E_m(\lambda) & (\xi_{ij})_{\substack{1\le i\le m\\ 0\le j\le m}} \\ O & I_{m+1} \end{pmatrix},$$

$$M_2 = \begin{pmatrix} I_m & O & O \\ (\zeta_j)_{1\le j\le m} & e(\nu) & (\zeta_{m+j})_{1\le j\le m} \\ O & O & I_m \end{pmatrix},$$

$$M_3 = \begin{pmatrix} I_{m+1} & O \\ (\eta_{ij})_{\substack{1\le i\le m\\ 1\le j\le m+1}} & E_m(\mu) \end{pmatrix},$$

$$N_1 = \begin{pmatrix} E_m(\lambda-\rho_1-\rho_2) & O \\ (\theta_{ij})_{\substack{0\le i\le m\\ 1\le j\le m}} & I_{m+1} \end{pmatrix},$$

$$N_2 = \begin{pmatrix} I_m & (\psi_i)_{1\le i\le m} & O \\ O & e(\nu-\rho_1-\rho_2) & O \\ O & (\psi_{m+i})_{1\le i\le m} & I_m \end{pmatrix},$$

$$N_3 = \begin{pmatrix} I_{m+1} & (\phi_{ij})_{\substack{1\le i\le m+1\\ 1\le j\le m}} \\ O & E_m(\mu-\rho_1-\rho_2) \end{pmatrix},$$

where

$$E_m(\lambda) = \begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_m) \end{pmatrix}, \qquad E_m(\lambda-\rho_1-\rho_2) = e(-\rho_1-\rho_2)E_m(\lambda),$$

$$E_m(\mu) = \begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_m) \end{pmatrix}, \qquad E_m(\mu-\rho_1-\rho_2) = e(-\rho_1-\rho_2)E_m(\mu),$$

$$\xi_{i0} = -e(\lambda_i+\nu-\rho_1-\rho_2)(e(\lambda_i)-e(\rho_1)) \cdot \prod_{k\in\{1,\dots,m\}\setminus\{i\}} \frac{1}{e(\rho_2-\lambda_i)-e(\rho_2-\lambda_k)} \quad (i=1,\dots,m),$$

$$\xi_{ij} = (e(\lambda_i)-e(\rho_1)) \cdot \prod_{k\in\{1,\dots,m\}\setminus\{i\}} \frac{e(\mu_j-\rho_1)-e(\rho_2-\lambda_k)}{e(\rho_2-\lambda_i)-e(\rho_2-\lambda_k)} \quad (i,j=1,\dots,m),$$

$$\zeta_j = -\prod_{\ell\in\{1,\dots,m\}} (e(\rho_2-\lambda_j)-e(\mu_\ell-\rho_1)) \quad (j=1,\dots,m),$$

$$\zeta_{m+j} = e(\rho_1) \cdot \prod_{k\in\{1,\dots,m\}} (e(\mu_j-\rho_1)-e(\rho_2-\lambda_k)) \quad (j=1,\dots,m),$$

$$\eta_{ij} = (e(\mu_i-\rho_1)-1) \cdot \prod_{\ell\in\{1,\dots,m\}\setminus\{i\}} \frac{e(\rho_2-\lambda_j)-e(\mu_\ell-\rho_1)}{e(\mu_i-\rho_1)-e(\mu_\ell-\rho_1)} \quad (i,j=1,\dots,m),$$

$$\eta_{i,m+1} = (e(\rho_1-\mu_i)-1) \cdot \prod_{\ell\in\{1,\dots,m\}\setminus\{i\}} \frac{1}{e(\mu_i-\rho_1)-e(\mu_\ell-\rho_1)} \quad (i=1,\dots,m),$$

and

$$\begin{aligned}
\theta_{0j} &= e(\lambda_j - \rho_1 - \rho_2) \cdot \zeta_j \quad (j = 1, \ldots, m),\\
\theta_{ij} &= e(-\mu_i) \cdot \eta_{ij} \quad (i, j = 1, \ldots, m),\\
\psi_i &= e(-\lambda_i) \cdot \xi_{i0} \quad (i = 1, \ldots, m),\\
\psi_{m+i} &= e(\nu - \rho_1 - \rho_2) \cdot \eta_{i,m+1} \quad (i = 1, \ldots, m),\\
\phi_{ij} &= e(\mu_j - \rho_1 - \rho_2) \cdot \xi_{ij} \quad (i, j = 1, \ldots, m),\\
\phi_{m+1,j} &= e(-\nu) \cdot \zeta_{m+j} \quad (j = 1, \ldots, m).
\end{aligned}$$

3.3. Monodromy group of system (IV).** Let $\lambda = (\lambda_1, \lambda_2)$, $\mu = (\mu_1, \mu_2)$, $\nu = (\nu_1, \nu_2)$, and $\rho = (\rho_1, \rho_2)$ be elements in $\mathbf{C}^2$ satisfying

$$\lambda_1 \neq \lambda_2, \quad \mu_1 \neq \mu_2, \quad \nu_1 \neq \nu_2, \quad \rho_1 \neq \rho_2,$$

and

$$\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + \nu_1 + \nu_2 = 4\rho_1 + 2\rho_2.$$

System $(\mathrm{IV}^{**})_{\lambda,\mu,\nu,\rho}$ (or simply (IV^{**})) is the system of total differential equations of rank 6

(3.5)

$$dZ = \Big\{(xI_6 - T_{\mathrm{IV}^*})^{-1} A_{\mathrm{IV}^*} dx + (A_{\mathrm{IV}^*} - (\rho_1 + \rho_2)I_6)(yI_6 - T_{\mathrm{IV}^*})^{-1} dy - A_{\mathrm{IV}^*} \frac{d(x-y)}{x-y}\Big\} Z$$

with

$$T_{\mathrm{IV}^*} = \begin{pmatrix} t_1 I_2 & & \\ & t_2 I_2 & \\ & & t_3 I_2 \end{pmatrix},$$

$$A_{\mathrm{IV}^*} = \begin{pmatrix}
\lambda_1 & & \alpha_{13} & \alpha_{14} & \alpha_{15} & \alpha_{16}\\
& \lambda_2 & \alpha_{23} & \alpha_{24} & \alpha_{25} & \alpha_{26}\\
\beta_{11} & \beta_{12} & \mu_1 & & \beta_{15} & \beta_{16}\\
\beta_{21} & \beta_{22} & & \mu_2 & \beta_{25} & \beta_{26}\\
\gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \nu_1 & \\
\gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & & \nu_2
\end{pmatrix},$$

where

$$\begin{aligned}
\alpha_{ij} &= \frac{\lambda_i - \rho_1}{\lambda_i - \lambda_{i'}} \cdot a_{ij} \qquad \text{for } i = 1, 2 \text{ with } \{i, i'\} = \{1, 2\}, \quad j = 3, 4, 5, 6,\\
\beta_{ij} &= \frac{\mu_i - \rho_1}{\mu_i - \mu_{i'}} \cdot b_{ij} \qquad \text{for } i = 1, 2 \text{ with } \{i, i'\} = \{1, 2\}, \quad j = 1, 2, 5, 6,\\
\gamma_{ij} &= \frac{\nu_i - \rho_1}{\nu_i - \nu_{i'}} \cdot c_{ij} \qquad \text{for } i = 1, 2 \text{ with } \{i, i'\} = \{1, 2\}, \quad j = 1, 2, 3, 4,
\end{aligned}$$

$$
\begin{aligned}
a_{13} &= \lambda_1 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2, & a_{14} &= \lambda_1 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2,\\
a_{15} &= \lambda_2 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2, & a_{16} &= \lambda_2 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2,\\
a_{23} &= \lambda_2 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2, & a_{24} &= \lambda_2 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2,\\
a_{25} &= \lambda_2 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2, & a_{26} &= \lambda_2 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2,\\
b_{11} &= \lambda_2 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2, & b_{12} &= \lambda_1 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2,\\
b_{15} &= \lambda_1 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2, & b_{16} &= \lambda_1 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2,\\
b_{21} &= \lambda_2 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2, & b_{22} &= \lambda_1 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2,\\
b_{25} &= \lambda_1 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2, & b_{26} &= \lambda_1 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2,\\
c_{11} &= \lambda_1 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2, & c_{12} &= \lambda_1 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2,\\
c_{13} &= \lambda_1 + \mu_2 + \nu_1 - 2\rho_1 - \rho_2, & c_{14} &= \lambda_1 + \mu_2 + \nu_2 - 2\rho_1 - \rho_2,\\
c_{21} &= \lambda_1 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2, & c_{22} &= \lambda_1 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2,\\
c_{23} &= \lambda_1 + \mu_1 + \nu_1 - 2\rho_1 - \rho_2, & c_{24} &= \lambda_1 + \mu_1 + \nu_2 - 2\rho_1 - \rho_2.
\end{aligned}
$$

The Jordan canonical form of A_{IV^*} is

$$
\begin{pmatrix} \rho_1 I_4 & \\ & \rho_2 I_2 \end{pmatrix}
$$

(see Haraoka [2, Thm. IV*]). Therefore, A_{IV^*} satisfies the relation

$$
(A_{\mathrm{IV}^*} - \rho_1 I_n)(A_{\mathrm{IV}^*} - \rho_2 I_n) = O,
$$

and hence the system (3.5) is completely integrable.

We assume the conditions

$$
\begin{aligned}
\lambda_1,\ \lambda_2,\ \lambda_1 - \lambda_2 &\notin \mathbf{Z},\\
\mu_1,\ \mu_2,\ \mu_1 - \mu_2 &\notin \mathbf{Z},\\
\nu_1,\ \nu_2,\ \nu_1 - \nu_2 &\notin \mathbf{Z},\\
\rho_1 - \rho_2 &\notin \mathbf{Z},\\
\rho_j &\notin \mathbf{Z}_{<0} \quad (j = 1, 2),
\end{aligned}
\tag{3.6}
$$

which are corresponding to (1.1). Then the system (3.5) has a fundamental set of solutions such as we state in section 1.1, which we denote by $\mathcal{Z}_{\mathrm{IV}^{**}}(x, y)$.

THEOREM 4. *We assume* (3.6) *and*

$$
\begin{aligned}
\lambda_i - \rho_1 &\notin \mathbf{Z} \quad (i = 1, 2),\\
\mu_i - \rho_1 &\notin \mathbf{Z} \quad (i = 1, 2),\\
\nu_i - \rho_1 &\notin \mathbf{Z} \quad (i = 1, 2),\\
\lambda_i + \mu_j + \nu_k - 2\rho_1 - \rho_2 &\notin \mathbf{Z} \quad (i, j, k = 1, 2).
\end{aligned}
$$

There is a diagonal matrix $D \in GL(n, \mathbf{C})$ such that the monodromy group of the *system* (3.5) *with respect to* $\mathcal{Z}_{\mathrm{IV}^{**}}(x,y)D$ *is generated by*

$$M_1 = \begin{pmatrix} e(\lambda_1) & & \xi_{11} & \xi_{12} & \xi_{13} & \xi_{14} \\ & e(\lambda_2) & \xi_{21} & \xi_{22} & \xi_{23} & \xi_{24} \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ \eta_{11} & \eta_{12} & e(\mu_1) & & \eta_{13} & \eta_{14} \\ \eta_{21} & \eta_{22} & & e(\mu_2) & \eta_{23} & \eta_{24} \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ \zeta_{11} & \zeta_{12} & \zeta_{13} & \zeta_{14} & e(\nu_1) & \\ \zeta_{21} & \zeta_{22} & \zeta_{23} & \zeta_{24} & & e(\nu_2) \end{pmatrix},$$

$$N_1 = \begin{pmatrix} e(\lambda_1-\rho_1-\rho_2) & & & & & \\ & e(\lambda_2-\rho_1-\rho_2) & & & & \\ \theta_{11} & \theta_{12} & 1 & & & \\ \theta_{21} & \theta_{22} & & 1 & & \\ \theta_{31} & \theta_{32} & & & 1 & \\ \theta_{41} & \theta_{42} & & & & 1 \end{pmatrix},$$

$$N_2 = \begin{pmatrix} 1 & & \phi_{11} & \phi_{12} & & \\ & 1 & \phi_{21} & \phi_{22} & & \\ & & e(\mu_1-\rho_1-\rho_2) & & & \\ & & & e(\mu_2-\rho_1-\rho_2) & & \\ & & \phi_{31} & \phi_{32} & 1 & \\ & & \phi_{41} & \phi_{42} & & 1 \end{pmatrix},$$

$$N_3 = \begin{pmatrix} 1 & & & & \psi_{11} & \psi_{12} \\ & 1 & & & \psi_{21} & \psi_{22} \\ & & 1 & & \psi_{31} & \psi_{32} \\ & & & 1 & \psi_{41} & \psi_{42} \\ & & & & e(\nu_1-\rho_1-\rho_2) & \\ & & & & & e(\nu_2-\rho_1-\rho_2) \end{pmatrix},$$

where

$$\xi_{ij} = \frac{e(\lambda_i)-e(\rho_1)}{e(\lambda_i)-e(\lambda_{i'})}\cdot x_{ij} \qquad \textit{for } i=1,2 \textit{ with } \{i,i'\}=\{1,2\}, \quad j=1,\ldots,4,$$
$$\eta_{ij} = \frac{e(\mu_i)-e(\rho_1)}{e(\mu_i)-e(\mu_{i'})}\cdot y_{ij} \qquad \textit{for } i=1,2 \textit{ with } \{i,i'\}=\{1,2\}, \quad j=1,\ldots,4,$$
$$\zeta_{ij} = \frac{e(\nu_i)-e(\rho_1)}{e(\nu_i)-e(\nu_{i'})}\cdot z_{ij} \qquad \textit{for } i=1,2 \textit{ with } \{i,i'\}=\{1,2\}, \quad j=1,\ldots,4,$$

$$\begin{aligned}
x_{11} &= \frac{[211]}{e(\nu_1+2\rho_1+\rho_2)}, & x_{12} &= \frac{[111]}{e(\nu_1+2\rho_1+\rho_2)},\\
x_{13} &= -\frac{[122]}{e(\lambda_1+\mu_2+\nu_1+3\rho_1+\rho_2)}, & x_{14} &= \frac{[111]}{e(\lambda_1+\mu_1+\nu_1+3\rho_1+\rho_2)},\\
x_{21} &= \frac{[221]}{e(\nu_1+2\rho_1+\rho_2)}, & x_{22} &= \frac{[121]}{e(\nu_1+2\rho_1+\rho_2)},\\
x_{23} &= -\frac{[212]}{e(\lambda_2+\mu_1+\nu_1+3\rho_1+\rho_2)}, & x_{24} &= \frac{[221]}{e(\lambda_2+\mu_1+\nu_1+3\rho_1+\rho_2)},\\
y_{11} &= \frac{[121]}{e(\rho_1)}, & y_{12} &= \frac{[111]}{e(\rho_1)},\\
y_{13} &= \frac{[212]}{e(\lambda_2+3\rho_1+\rho_2)}, & y_{14} &= -\frac{[222]}{e(\lambda_2+3\rho_1+\rho_2)},\\
y_{21} &= \frac{[221]}{e(\rho_1)}, & y_{22} &= \frac{[211]}{e(\rho_1)},\\
y_{23} &= -\frac{[211]}{e(\lambda_1+\lambda_2+\mu_1+\nu_1+\rho_1)}, & y_{24} &= \frac{[221]}{e(\lambda_1+\lambda_2+\mu_1+\nu_1+\rho_1)},\\
z_{11} &= -e(\lambda_1+\nu_1)[221], & z_{12} &= -e(\lambda_2+\nu_1)[111],\\
z_{13} &= [221], & z_{14} &= [111],\\
z_{21} &= -\frac{e(2\rho_1+\rho_2)[121]}{e(\mu_2)}, & z_{22} &= -\frac{e(2\rho_1+\rho_2)[211]}{e(\mu_2)},\\
z_{23} &= -[211], & z_{24} &= \frac{e(2\rho_1+\rho_2)[212]}{e(\lambda_2+\mu_2+\nu_2)},
\end{aligned}$$

and

$$\begin{aligned}
\theta_{ij} &= e(\lambda_j-\rho_1-\rho_2)\cdot\eta_{ij}, & \theta_{2+i,j} &= e(-\nu_i)\cdot\zeta_{ij},\\
\phi_{ij} &= e(-\lambda_i)\cdot\xi_{ij}, & \phi_{2+i,j} &= e(\mu_j-\rho_1-\rho_2)\cdot\zeta_{i,2+j},\\
\psi_{ij} &= e(\nu_j-\rho_1-\rho_2)\cdot\xi_{i,2+j}, & \psi_{2+i,j} &= e(-\mu_i)\cdot\eta_{i,2+j}
\end{aligned}$$

for $i,j=1,2$. Here we have set

$$[ijk]=e(\lambda_i+\mu_j+\nu_k)-e(2\rho_1+\rho_2)$$

for $i,j,k=1,2$.

3.4. Proof of Theorems 2, 3, and 4. The elements of M_i $(i=1,2,3)$ have already been determined by Haraoka [3, Thms. 7, 9, and 11]. Applying the following lemma to the M_i's, we can easily evaluate the elements of N_j $(j=1,2,3)$.

LEMMA 2. *Let M_i $(i=1,2,3)$ and N_j $(j=1,2,3)$ be matrices of the form*

$$M_1=\begin{pmatrix} M_{11} & M_{12} & M_{13}\\ & I_{n_2} & \\ & & I_{n_3}\end{pmatrix}, M_2=\begin{pmatrix} I_{n_1} & & \\ M_{21} & M_{22} & M_{23}\\ & & I_{n_3}\end{pmatrix}, M_3=\begin{pmatrix} I_{n_1} & & \\ & I_{n_2} & \\ M_{31} & M_{32} & M_{33}\end{pmatrix},$$

where M_{ij} is an $n_i\times n_j$ matrix, and

$$N_1=\begin{pmatrix} N_{11} & & \\ N_{21} & I_{n_2} & \\ N_{31} & & I_{n_3}\end{pmatrix}, N_2=\begin{pmatrix} I_{n_1} & N_{12} & \\ & N_{22} & \\ & N_{32} & I_{n_3}\end{pmatrix}, N_3=\begin{pmatrix} I_{n_1} & & N_{13}\\ & I_{n_2} & N_{23}\\ & & N_{33}\end{pmatrix},$$

where N_{ij} is an $n_i \times n_j$ matrix. Let ε be an element in $\mathbf{C} \setminus \{0\}$. Suppose that M_i $(i = 1, 2, 3)$ and N_j $(j = 1, 2, 3)$ are invertible and satisfy

$$N_1N_2N_3 = \varepsilon M_3M_2M_1 \tag{3.7}$$

and

$$N_jM_i = M_iN_j \tag{3.8}$$

for $i, j = 1, 2, 3$ with $i \neq j$. Then we have

$$N_{11} = \varepsilon M_{11}, \qquad N_{21} = \varepsilon M_{21}M_{11}, \quad N_{31} = M_{33}^{-1}M_{31}, \tag{3.9}$$

$$N_{12} = M_{11}^{-1}M_{12}, \quad N_{22} = \varepsilon M_{22}, \qquad N_{32} = \varepsilon M_{32}M_{22}, \tag{3.10}$$

and

$$N_{13} = \varepsilon M_{13}M_{33}, \quad N_{23} = M_{22}^{-1}M_{23}, \quad N_{33} = \varepsilon M_{33}. \tag{3.11}$$

Proof. We prove (3.10). By direct calculation, we obtain

$$N_1N_2N_3 = \begin{pmatrix} N_{11} & * & * \\ N_{21} & * & * \\ N_{31} & * & * \end{pmatrix}$$

and

(3.12)
$$\varepsilon M_3M_2M_1 = \varepsilon \begin{pmatrix} M_{11} & M_{12} & * \\ M_{21}M_{11} & M_{21}M_{12} + M_{22} & * \\ (M_{31} + M_{32}M_{21})M_{11} & (M_{31} + M_{32}M_{21})M_{12} + M_{32}M_{22} & * \end{pmatrix}.$$

From these with the relation (3.7) we temporarily obtain

$$N_{11} = \varepsilon M_{11}, \quad N_{21} = \varepsilon M_{21}M_{11}, \quad N_{31} = \varepsilon(M_{31} + M_{32}M_{21})M_{11},$$

and hence

$$N_1^{-1} = \begin{pmatrix} \varepsilon^{-1}M_{11}^{-1} & & \\ -M_{21} & I_{n_2} & \\ -(M_{31} + M_{32}M_{21}) & & I_{n_3} \end{pmatrix}. \tag{3.13}$$

From (3.12) and (3.13) we obtain

$$\varepsilon N_1^{-1}M_3M_2M_1 = \begin{pmatrix} I_{n_1} & M_{11}^{-1}M_{12} & * \\ & \varepsilon M_{22} & * \\ & \varepsilon M_{32}M_{22} & * \end{pmatrix}.$$

On the other hand, by the relation (3.7) multiplied by N_1^{-1} from the left we have

$$\varepsilon N_1^{-1}M_3M_2M_1 = N_2N_3 = \begin{pmatrix} I_{n_1} & N_{12} & * \\ & N_{22} & * \\ & N_{32} & * \end{pmatrix}.$$

Comparing these expressions, we obtain (3.10).

Starting with the relations

$$N_2N_3N_1 = \varepsilon M_1M_3M_2 \quad \text{and} \quad N_3N_1N_2 = \varepsilon M_2M_1M_3,$$

which are derived from (3.7) with (3.8), we can prove (3.11) and (3.9), respectively, by the same consideration as above. Thus the proof is complete. $\square$

REFERENCES

[1] Y. Haraoka, *Finite monodromy of Pochhammer equation*, Ann. Inst. Fourier, 44 (1994), pp. 767–810.

[2] Y. Haraoka, *Canonical forms of differential equations free from accessory parameters*, SIAM J. Math. Anal., 25 (1994), pp. 1203–1226.

[3] Y. Haraoka, *Monodromy representations of systems of differential equations free from accessory parameters*, SIAM J. Math. Anal., 25 (1994), pp. 1595–1621.

[4] K. Okubo, *On the group of Fuchsian equations*, Seminar Reports, Tokyo Metropolitan Univ., 1987.

[5] K. Okubo, K. Takano, and S. Yoshida, *A connection problem for the generalized hypergeometric equation*, Funkcial. Ekvac., 31 (1988), pp. 483–495.

[6] T. Sasai, *On a monodromy group and irreducibility conditions of a fourth order Fuchsian differential system of Okubo type*, J. Reine Angew. Math., 299/300 (1978), pp. 38–50.

[7] T. Sasai and S. Tsuchiya, *On a class of even order Fuchsian equations of Okubo type*, Funkcial. Ekvac., 35 (1992), pp. 505–514.

[8] K. Takano and E. Bannai, *A global study of Jordan–Pochhammer differential equations*, Funkcial. Ekvac., 19 (1976), pp. 85–99.

[9] T. Yokoyama, *A system of total differential equations of two variables and its monodromy group*, Funkcial. Ekvac., 35 (1992), pp. 65–93.

[10] T. Yokoyama, *On an irreducibility condition for hypergeometric systems*, Funkcial. Ekvac., 38 (1995), pp. 11–19.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1248–1257, September 1997

015

ASYMPTOTICS OF THE ZEROS OF RELATIVISTIC HERMITE POLYNOMIALS*

MATTHEW HE[†], K. PAN[‡], AND PAOLO E. RICCI[§]

Abstract. The relativistic Hermite polynomial (RHP) is a class of orthogonal polynomials associated with varying weights. We study the asymptotics of the zeros of the RHP when both degree n of polynomials and relativistic parameter N approach infinity.

Key words. Hermite polynomials, orthogonal polynomials

AMS subject classifications. 41A, 41C

PII. S003614109529642X

1. Introduction.

Relativistic Hermite polynomials (RHPs) $\{H_n^{(N)}(x)\}_{n=0}^{\infty}$ were introduced in [1] in connection with the wave functions of the quantum relativistic harmonic oscillator. It was shown in [1] that the RHP satisfies the second-order differential equation

$$\left(1+\frac{x^2}{N}\right)y_n'' - \frac{2}{N}(N+n-1)xy_n' + \frac{n}{N}(2N+n-1)y_n = 0. \tag{1.1}$$

Equation (1.1) is a particular case of a second-order hypergeometric-type equation [9]

$$\sigma(x)y'' + \tau(x)y' + \lambda y = 0, \tag{1.2}$$

where

$$\sigma(x) = \left(1+\frac{x^2}{N}\right),$$
$$\tau = -\frac{2}{N}(N+n-1)x,$$
$$\lambda = \frac{n}{N}(2N+n-1).$$

It is easy to verify that the following relation holds:

$$\lambda = -n\tau' - \frac{1}{2}n(n-1)\sigma''.$$

By solving the equation

$$[\sigma(x)\rho_n(x;N)]' = \tau(x)\rho_n(x;N),$$

*Received by the editors December 18, 1995; accepted for publication April 17, 1996.
http://www.siam.org/journals/sima/28-5/29642.hmtl

[†] Department of Mathematics, Nova Southeastern University, Ft. Lauderdale, FL 33314 (hem@polaris.ncs.nova.edu).

[‡] Department of Mathematics and Computer Science, Barry University, Miami Shores, FL 33161 (pan@dominic.barry.edu).

[§] Dipartimento di Mathematica "Guido Castelnuovo," Universita degli Studi di Roma "La Sapienza," 00185 Rome, Italy (riccip@sci.uniroma1.it).

one can find the symmetric factor or weight function

$$\rho_n(x;N) = \left(1+\frac{x^2}{N}\right)^{-(N+n)}, \quad N > \frac{1}{2}, \quad n = 0,1,2,\ldots,\ x \in (-\infty,\infty). \tag{1.3}$$

Using this weight function, the following orthogonality of the RHP was established in [1]:

$$\int_{-\infty}^{\infty} x^k H_n^{(N)}(x)\rho_n(x;N)dx = 0, \qquad k = 0,1,\ldots,n-1. \tag{1.4}$$

That is, $\{H_n^{(N)}(x)\}_{n=0}^{\infty}$ is a class of orthogonal polynomials with respect to a sequence of varying weight functions $\rho_n(x)$. Clearly,

$$\lim_{N\to\infty} \rho_n(x;N) = e^{-x^2}$$

and

$$\lim_{N\to\infty} H_n^{(N)}(x) = H_n(x).$$

So the relativistic Hermite polynomials become classical Hermite polynomials when the relativistic parameter $N \to \infty$.

The distributions of zeros of RHP were studied in [2]. An analytic approximation for the distribution was derived within the framework of the WKB approximation.

The asymptotics of orthogonal polynomials with respect to varying weights are closely related to constrained or weighted polynomial approximation. Logarithmic potential has been extensively used in investigating such asymptotics. We study the asymptotics of the zeros of RHP when both n and N approach ∞ by using the potential-theoretic method.

The paper is organized as follows: in order to state our main results, we shall introduce some basics from potential theory in section 2. Applying a general result from potential theory developed in [11] to our relativistic weight function, we determine the support of the equilibrium measure explicitly in section 3. In section 4, we give an explicit formula for the density function of the equilibrium measure. The asymptotics of the zeros of the RHP when both n and N approach ∞ are determined in section 5.

2. Basics of potential theory. We shall use logarithmic potentials of Borel measures. If μ is a finite Borel measure with compact support, then its logarithmic potential is defined as its convolution with the logarithmic kernel:

$$U^{\mu}(z) = \int \log\frac{1}{|z-t|}d\mu(t).$$

Let E be a closed subset of the real number line. A weight function w on E is said to be admissible if it satisfies the following three conditions:

(i) w is continuous;

(ii) $\text{Cap}\{x \in E \mid w(x) > 0\} > 0$;

(iii) $Z := \{x \in E : w(x) = 0\}$ has capacity zero; and

(iv) if E is unbounded, then $|x|w(x) \to 0$ as $|x| \to \infty$, $x \in E$.

We say that w is strongly admissible if

(i) w^q is admissible for every q, $0 < q \leq 1$;

(ii) E is regular, i.e., for all k large, $E \cap [-k,k]$ is regular with respect to the Dirichlet problem for its complement on the Riemann sphere, and

(iii) $E \backslash Z$ is interval-like.

We define $Q = Q_w$ by

$$w(x) = \exp(-Q(x)). \tag{2.1}$$

Then $Q : E \to (-\infty, \infty]$ is continuous everywhere where w is positive, i.e., Q is finite.

Let $\mathcal{M}(E)$ be the set of all positive unit Borel measures μ with support $S(\mu) := \operatorname{supp}(\mu) \subset E$ and define the weighted energy integral

$$I_w[\mu] = \iint \log \frac{1}{|z-t|w(z)w(t)} d\mu(z)d\mu(t).$$

Let

$$V_w(E) = \inf\{I_w[\mu] \mid \mu \in \mathcal{M}(E)\}.$$

Then the following properties are true (cf. [12], [10]).

(i) $V_w(E)$ is finite.

(ii) There exists a unique $\mu_E \in \mathcal{M}(E)$ such that

$$I_w(\mu_w) = V_w(E).$$

Moreover, μ_w has finite logarithmic energy.

(iii) $S(\mu_w)$ is a compact subset of E.

(iv) The inequality

$$U^{\mu_w}(z) + Q(z) \geq F_w, \quad z \in E. \tag{2.2}$$

(v) The equality

$$U^{\mu_w}(z) + Q(z) = F_w, \quad z \in S(\mu_w). \tag{2.3}$$

The measure μ_w is called the equilibrium or extremal measure in the presence of an external field, and

$$F_w = V_w(E) - \int Q d\mu_w. \tag{2.4}$$

In order to state our applications to polynomial extremal problems, we define

$$E_{n,p}(w) := \inf\{\|[w(x)]^n[x^n - P(x)]\|_{E,p} : P \in \mathcal{P}_{n-1}\},$$

where $\mathcal{P}_n$ is the set of all polynomials with degrees $\leq n$ and

$$\|f\|_{E,p} := \left(\int_E |f|^p dx\right)^{1/p},$$

$n = 1, 2, \ldots, 0 < p \leq \infty$. The extremal polynomials $T_n(x; w, p) = x^n + \cdots \in \mathcal{P}_n$ are defined by the property

$$E_{n,p}(x) = \|[w(x)]^n T_n(x; w, p)\|_{E,p}.$$

Finally, in this section, we state the following two lemmas.

Lemma 2.1 (see [6]). *Let w be strongly admissible and $0 < p \leq \infty$. Let $\{t_{n,k}\}_{k=1}^{n}$ be the zeros of $T_n(x; w, p)$. Then there exists a closed bounded interval I containing $S(\mu_w)$ and all the zeros of $T_n(x; w, p)$. Moreover,*

$$\lim_{n\to\infty} |T_n(x; w, p)|^{1/n} = \exp\left[\int \log|z - t| d\mu_w(t)\right]$$

uniformly on every compact set of the complex plane disjoint from I,

$$\lim_{n\to\infty} [E_{n,p}]^{1/n} = \exp(F_w),$$

and

$$\lim_{n\to\infty} \mu_n = \mu_w$$

in the weak-star topology, where

$$\mu_n(B) := \frac{1}{n}\#\{k : t_{n,k} \in B\}, \qquad n = 1, 2, \ldots,$$

for any Borel set B.

Lemma 2.2 (see [6]). *Let w be strongly admissible and $0 < p \leq \infty$. Suppose that $I \subset \mathbf{R}$ is a closed bounded interval containing $S(\mu_w)$. Let $\{v_{n,k}\}_{k=1}^{n}$ be a triangular scheme of points lying in I. With this scheme, let $q_n(x) = \prod_{k=1}^{n}(x - v_{n,k})$. Assume that for some p $(0 < p \leq \infty)$,*

$$\lim_{n\to\infty} \|w^n q_n\|_{E,p}^{1/n} \leq \exp(F_w).$$

Then

$$\lim_{n\to\infty} |q_n(x)|^{1/n} = \exp\left[\int \log|z - t| d\mu_w(t)\right]$$

uniformly on every compact set of the complex plane disjoint from I, and

$$\lim_{n\to\infty} \mu_n = \mu_w$$

in the weak-star topology, where

$$\mu_n(B) := \frac{1}{n}|\{k : v_{n,k} \in B\}|, \qquad n = 1, 2, \ldots,$$

for any Borel set B.

3. Support of equilibrium measure. A fundamental theorem [5] in weighted polynomial approximation asserts that every weighted polynomial $\{w^n(x)p_n(x)\}$ must assume its maximum modulus on $S(\mu)$, i.e.,

$$\|w^n(x)p_n(x)\|_E = \|w^n(x)p_n(x)\|_{S(\mu)}, \tag{3.1}$$

where $S(\mu)$ is the support of the equilibrium measure of the set E, and $\|\cdot\|_E$ is the sup norm.

In this section we determine explicitly the support of the equilibrium measure $S(\mu)$ for the weight function $\rho_n(x;N)$. To find $S(\mu)$, we shall need to directly maximize the following F-functional [5]:

$$F(a,b) = \log\left(\frac{(b-a)}{4}\right) - \frac{1}{\pi}\int_a^b \frac{Q(x)}{\sqrt{(x-a)(b-x)}}\,dx.$$

We define $w_n(x) = \rho_n^{\frac{1}{2n}}(x;N)$. Then we have

$$Q_n(x) = \log\frac{1}{w_n(x)} = \left(\frac{1}{2}+\frac{N}{2n}\right)\log\left(1+\frac{x^2}{N}\right).$$

Since $Q_n(x)$ is an even function, the F-functional can be written as follows:

$$F(a) := f(-a,a) = \log a - \frac{1}{\pi}\left(1+\frac{N}{n}\right)\int_0^a \frac{\log\left(1+\frac{t^2}{N}\right)}{\sqrt{a^2-t^2}}\,dt - \log 2.$$

By an elementary integral formula [3],

$$\int_0^1 \frac{\log(1+bx^2)}{\sqrt{1-x^2}}\,ds = \pi\log\frac{1+\sqrt{1+b}}{2}.$$

We have

$$F(a) = \log\frac{a}{2} - \left(1+\frac{N}{n}\right)\log\frac{1+\sqrt{1+\frac{a^2}{N}}}{2}.$$

It is now elementary to check that the choice of $a = a_n$, which maximizes $F(-a,a)$, is given by

$$a_n = \sqrt{\frac{n(n+2N)}{N}}. \tag{3.2}$$

Therefore, we have determined the support $[-a_n, a_n]$ of equilibrium measure corresponding to varying weight $\rho_n(x;N)$.

Furthermore, we can determine the constant F_{w_n},

$$F_{w_n} = \log\frac{a_n}{2} - \left(1+\frac{N}{n}\right)\log\frac{1+\sqrt{1+\frac{a_n^2}{N}}}{2}.$$

We note that

$$\lim_{N\to\infty} a_n = \sqrt{2n},$$

$$\lim_{n\to\infty} F_{w_n} = -\frac{1}{2}\log N,$$

which concides with the results of [8]. We remark here that, although we use a potential-theoretic approach similar to the one used in [8], our approach is more direct. We shall continue our investigation along the same direction to determine the equilibrium measure.

4. Equilibrium measure. In section 3, we determined the support $S(\mu_n) = [-a_n, a_n]$ of equilibrium measure μ_n associated with varying weight $w_n(x)$. In this section, we apply a general formula [12, p. 53] for the density function of the equilibrium measure to our weight function $w_n(x)$ and find the following theorem.

THEOREM 4.1.

$$d\mu_n(t) = g_n(t)dt = \frac{N}{n\pi}\,\frac{\sqrt{a_n^2 - t^2}}{N + t^2}\,dt, \qquad t \in S(\mu_n). \tag{4.1}$$

Proof. It was shown in [12, Lem. 5.1] that the integral equation

$$\int_{-1}^{1} \log\frac{1}{|x-t|}g(t)dt = -Q(x) + C,$$

where C is some constant, has a solution $g(t)$ of the form

$$g(t) = \frac{2}{\pi^2}\sqrt{1-t^2}\int_0^1 \frac{sQ'(s) - tQ'(t)}{(1-s^2)^{1/2}(s^2-t^2)}\,ds + \frac{D_1}{\sqrt{1-t^2}}, \tag{4.2}$$

where

$$D_1 = \frac{1}{\pi} - \frac{1}{\pi^2}\int_{-1}^{1}\frac{sQ'(s)}{\sqrt{1-s^2}}\,ds. \tag{4.3}$$

$g(t)$ is even and has total integral 1 over $[-1, 1]$. Apply (4.2) and (4.3) to

$$Q_n(a_n x) = \left(\frac{1}{2} + \frac{N}{2n}\right)\log\left(1 + \frac{a_n^2 x^2}{N}\right),$$

and we get

$$g_n(t) = \frac{N}{n\pi}\,\frac{\sqrt{a_n^2 - t^2}}{N + t^2}.$$

We note that

$$\lim_{n\to\infty} g_n(t) = \frac{\sqrt{N}}{\pi(N+t^2)}, \qquad t \in (-\infty, \infty).\square$$

5. Asymptotics of zeros. In this section, we study the zeros distribution of $H_n^{(N)}(x)$ for n, $N \to \infty$. The following lemma tells us that the support of $H_n^{(N)}(x)$ "lives" also in some compact set in L_2.

LEMMA 5.3. *For $w(x) = (1 + x^2/N)^{(-N-n)/2}$, there is a positive constant A independent of n, N, such that, for $p \in \mathcal{P}_n$,*

$$\|w(x)p(x)\|_{(-\infty,\infty),2} \le 2\|w(x)p(x)\|_{[-Aa_n, Aa_n],2}.$$

Proof. The proof can be found in [4] for the fixed weight. Here we have a varying weight, so we may proceed exactly as in Theorem 5.2 in [4] to get the lemma. □

The next theorem will discuss the location of the zeros of $H_n^{(N)}(x)$.

Theorem 5.2. *For $w(x)=(1+x^2/N)^{(-N-n)/2}$, there is a positive constant D independent of n, N, such that all the zeros of $H_n^{(N)}(x)$ lie in $[-Da_n, Da_n]$.*

Proof. Let $X_{n,N}$ denote the largest zero of $H_n^{(N)}(x)$. Suppose now that $\forall A>0$ there exist n, N such that $X_{n,N}>Aa_n$. Let

$$t_{n,N}(x) := \frac{x-Aa_n}{x-X_{n,N}} H_n^{(N)}(x).$$

Then, for $x\in[-Aa_n, Aa_n]$,

$$|t_{n,N}(x)| \le \frac{2Aa_n}{X_{n,N}-Aa_n}|H_n^{(N)}(x)|.$$

Hence, from the lemma above, we have

$$\begin{aligned}
&\|w(x)t_{n,N}(x)\|_{(-\infty,\infty),2}\\
&\quad\le\ 2\|w(x)t_{n,N}(x)\|_{[-Aa_n,Aa_n],2}\\
&\quad\le\ 2\left(\frac{2Aa_n}{X_{n,N}-Aa_{n,N}}\right)\|w(x)H_n^{(N)}(x)\|_{[-Aa_n,Aa_n],2}\\
&\quad\le\ 2\left(\frac{2Aa_n}{X_{n,N}-Aa_{n,N}}\right)\|w(x)H_n^{(N)}(x)\|_{(-\infty,\infty),2}.
\end{aligned}$$

Thus, since $H_n^{(N)}(x)$ is extremal, the inequality implies that $1\le 4Aa_n/(X_{n,N}-Aa_n)$, that is, $X_{n,N}\le 5Aa_n$. □

Now, we consider the case in which both n and N converge to ∞ with the same rate.

Theorem 5.3. *Let $N=\lambda n$ and a_n be as in (3.2), where λ is a fixed number. Then*

$$\lim_{n\to\infty}|H_n^{(N)}(a_nx)|^{\frac{1}{n}} = \exp\left[\int_{-1}^{1}\log|z-t|d\mu(t)\right]$$

locally uniformly in $\mathbf{C}\backslash[-1,1]$, where

$$d\mu(t) = \frac{\lambda(1+2\lambda)}{\pi}\quad\frac{\sqrt{1-t^2}}{\lambda^2+(1+2\lambda)t^2}\,dt.$$

Furthermore, let $\{t_{n,k}\}_{k=1}^{n}$ be the zeros of $H_n^{(N)}(a_nx)$ and B be a Borel set. Define

$$\mu_n := \frac{1}{n}|\{k : t_{n,k}\in B\}|,\qquad n=1,2,\ldots;$$

then

$$\lim_{n\to\infty}\mu_n = \mu$$

in the weak-star topology.

Proof. Let

$$w(x) = \left(1+\frac{(1+2\lambda)}{\lambda^2}x^2\right)^{-\frac{1+\lambda}{2}}.$$

Let $t_n(x) = x^n + \cdots$ be the extremal polynomial for the sup norm on $[-1, 1]$,

$$\|w(x)^n t_n(x)\| = \inf_{p_n = x^n + \cdots} \|w(x)^n p_n(x)\|,$$

and $T_n(x) = a_n^n t_n(\frac{x}{a_n})$. Notice that $a_n = \sqrt{(1 + 2\lambda)n/\lambda}$; then

$$\begin{aligned}
&\|w(x)^n H_n^{(N)}(a_n x)\|_{[-1,1],2} \\
&\quad = \left\| \left(1 + \frac{(1+2\lambda)}{\lambda^2} x^2\right)^{-\frac{1+\lambda}{2}n} H_n^{(N)}(a_n x) \right\|_{[-1,1],2} \\
&\quad = \left\| \left(1 + \frac{(1+2\lambda)}{\lambda^2} \, \frac{x^2}{a_n^2}\right)^{-\frac{1+\lambda}{2}n} H_n^{(N)}(x) \right\|_{[-a_n,a_n],2} \\
&\quad \le \left\| \left(1 + \frac{x^2}{\lambda n}\right)^{-\frac{1+\lambda}{2}n} H_n^{(N)}(x) \right\|_{(-\infty,\infty),2} \\
&\quad \le \left\| \left(1 + \frac{x^2}{\lambda n}\right)^{-\frac{1+\lambda}{2}n} T_n(x) \right\|_{(-\infty,\infty),2} \\
&\quad \le \left\| \left(1 + \frac{x^2}{\lambda n}\right)^{-\frac{1+\lambda}{2}n} T_n(x) \right\|_{-(\infty,\infty),\infty} \\
&\quad = \left\| \left(1 + \frac{x^2}{\lambda n}\right)^{-\frac{1+\lambda}{2}n} T_n(x) \right\|_{[-a_n,a_n],\infty} \\
&\quad = \left\| \left(1 + \frac{(1+2\lambda)}{\lambda^2} x^2\right)^{-\frac{1+\lambda}{2}n} T_n(a_n x) \right\|_{[-1,1],\infty} \\
&\quad = \left\| \left(1 + \frac{(1+2\lambda)}{\lambda^2} x^2\right)^{-\frac{1+\lambda}{2}n} a_n^n t_n(x) \right\|_{[-1,1],\infty}.
\end{aligned}$$

Thus we have

$$\left\| [w(x)]^n \frac{H_n(a_n x)}{a_n^n} \right\|_{[-1,1],2}^{\frac{1}{n}} \le \left\| \left(1 + \frac{(1+2\lambda)}{\lambda^2} \, x^2\right)^{-\frac{1+\lambda}{2}n} t_n(x) \right\|_{[-1,1],\infty}^{1/n}. \tag{5.1}$$

For the weight w, notice that $t_n(x)$ is the extremal for w, and from Lemma 1, we have

$$\limsup_{n\to\infty} \left\| w(x)^n \frac{H_n(a_n x)}{a_n^n} \right\|_{[-1,1],2}^{\frac{1}{n}} \le \exp(F_w).$$

Notice that $\frac{H_n(a_n x)}{a_n^n} = x^n + \cdots$ and the zeros of $H_n^{(N)}(a_n x)$ lie in $[-D, D]$. From Lemma 2, we have the proof of the theorem. □

THEOREM 5.4. *Let $N = \lambda_n n$ and a_n be as in* (3.2). *If $\lambda_n \to \infty$, then*

$$\lim_{n\to\infty} |H_n^{(N)}(a_n x)|^{\frac{1}{n}} = \exp\left[\int_{-1}^{1} \log|z - t| d\nu(t)\right]$$

locally uniformly in $\mathbf{C}\backslash[-1, 1]$, where

$$d\nu(t) = \frac{2}{\pi} \sqrt{1 - t^2} \, dt.$$

Furthermore, let the sequence of unit measures $\{\nu_n\}_{n=1}^{\infty}$ *be*

$$\nu_n := \frac{1}{n}|\{k : t_{n,k} \in B\}|, \qquad n = 1, 2, \ldots,$$

where B *is a Borel set and* $\{t_{n,k}\}_{k=1}^{n}$ *are the zeros of* $H_n^{(N)}(a_n x)$*. If* $\lambda_n \to \infty$*, then*

$$\lim_{n-\infty} \nu_n = \nu$$

in the weak-star topology.

Proof. In the proof of (5.1), it is easy to see that for any $p_n(x) = x^n + \cdots \in \mathcal{P}_n$, we have

$$\left\|\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n} \frac{H_n(a_n x)}{a_n^n}\right\|_{[-1,1],2}$$

(5.2)

$$\leq \left\|\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n} p_n(x)\right\|_{[-1,1],\infty}.$$

Here, we consider $w(x) = e^{-x^2}$ on $[-1,1]$; the equilibrium measure is $d\nu = \frac{2}{\pi}\sqrt{1-t^2}\,dt$ [7]. Choose $p_n(x) = T_n(x; w, \infty)$; from (5.2), we have

$$\begin{aligned}
&\left\|[w(x)]^n \frac{H_n^{(N)}(a_n x)}{a_n^n}\right\|_{[-1,1],2}\\
&= \left\|e^{-nx^2} \frac{H_n^{(N)}(a_n x)}{a_n^n}\right\|_{[-1,1],2}\\
&\leq \left\|e^{-nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],2} \left\|\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n} \frac{H_n^{(N)}(a_n x)}{a_n^n}\right\|_{[-1,1],2}\\
&\leq \left\|e^{-nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],2} \left\|\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n} T_n(x; w, \infty)\right\|_{[-1,1],\infty}\\
&\leq \left\|e^{-nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],2} \left\|e^{nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],\infty}\\
&\times \left\|e^{-nx^2} T_n(x; w, \infty)\right\|_{[-1,1],\infty}.
\end{aligned}$$

Notice that, as $\lambda_n \to \infty$,

$$\lim_{n\to\infty} \left\|e^{-nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],2}^{1/n} = 1$$

and

$$\lim_{n\to\infty} \left\|e^{nx^2}\left(1 + \frac{(1+2\lambda_n)}{\lambda_n^2}x^2\right)^{-\frac{1+\lambda_n}{2}n}\right\|_{[-1,1],\infty}^{1/n} = 1;$$

then

$$\lim_{n\to\infty}\left\|[w(x)]^n\frac{H_n^{(N)}(a_nx)}{a_n^n}\right\|_{[-1,1],2}^{1/n}\le e^{F_w}.$$

From Lemma 2, this completes the proof of the theorem. □

For the case when N is fixed and $n\to\infty$, see [8].

REFERENCES

[1] V. Aldaya, J. Bisquert, and J. Navarro-Salas, *The quantum relativistic harmonic oscillator: Generalized Hermite polynomials*, Phys. Lett. A, 156 (1991), pp. 381–385.

[2] J. S. Dehesa, J. Torres, and A. Zarzo, *On a new set of polynomials representing the wave functions of the quantum relativistic harmonic oscillator*, J. Phys. A: Math. Gen., to appear.

[3] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[4] D. S. Lubinsky, *Strong Asymptotics for Extremal Errors and Polynomials Associated with Erdos-Type Weights*, Pitman Research Notes, Vol. 202, Longman, Harlow, UK, 1988.

[5] H. N. Mhaskar and E. B. Saff, *Where does the sup norm of a weighted polynomial live?*, Constr. Approx., 1 (1985), pp. 71–91.

[6] H. N. Mhaskar and E. B. Saff, *Where does the L^p-norm of a weighted polynomial live?*, Trans. Amer. Math. Soc., 303 (1987), pp. 109–124.

[7] H. N. Mhaskar and E. B. Saff, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203–234.

[8] A. Zarzo and A. Martinez, *The quantum relativistic harmonic oscillator: Spectrum of zeros of its wave functions*, J. Math. Phys., 34 (1993), pp. 2926–2935.

[9] A. F. Nikiforov and V. B. Uvarov, *Special Functions of Mathematical Physics*, Birkhäuser, Basel, 1988.

[10] E. B. Saff and V. Totik, *Logarithmic Potential with External Fields*, Springer-Verlag, Berlin, 1995.

[11] H. Stahl and V. Totik, *General Orthogonal Polynomials*, Encyclopedia of Mathematics 43, Cambridge University Press, New York, 1992.

[12] V. Totik, *Weighted Approximation with Varying Weight*, Lecture Notes in Mathematics 1569, Springer-Verlag, New York, 1994.

SIAM J. Math. Anal.
Vol. 28, No. 5, pp. 1258–1263, September 1997

ORTHOGONALITY OF CARDINAL B-SPLINES IN WEIGHTED SOBOLEV SPACES *

ULRICH REIF[†]

Abstract. The cardinal B-splines $B_{j,n}, j \in \mathbb{Z}$, of order n form an orthonormal sequence in the Sobolev space $H^{n-1,2}(\mathbb{R})$ endowed with the norm $\|f\|^2_{\omega(n)} := \sum_{\mu=0}^{n-1} \omega_\mu(n) \|\partial^\mu f\|^2$ for certain positive weights $\omega_\mu(n)$. These weights are specified explicitly. Further, an application to approximation theory is discussed.

Key words. B-spline, Sobolev space, orthogonality, approximation

AMS subject classifications. 65D07, 41A15, 41A29

PII. S0036141096296851

1. Preliminaries.

In this section we shall briefly introduce some basic concepts from B-spline theory and functional analysis; see, e.g., [4], [3], [1], and [5] for an introduction to these topics.

For $m \in \mathbb{N}$ denote by $\langle \cdot, \cdot \rangle$, $\langle \cdot, \cdot \rangle_m$ the inner products and by $\| \cdot \|$, $\| \cdot \|_m$ the norms of the Hilbert spaces $L^2(\mathbb{R})$ and $H^{m,2}(\mathbb{R})$, respectively. Let $\omega := [\omega_0, \dots, \omega_m]$ be a vector of positive weights, and define the *weighted Sobolev space* $H^{m,2}_\omega(\mathbb{R})$ by providing $H^{m,2}(\mathbb{R})$ with the inner product

$$(f,g)_\omega := \sum_{\mu=0}^{m} \omega_\mu \langle \partial^\mu f, \partial^\mu g \rangle \; . \tag{1.1}$$

Evidently, the induced norm $\| \cdot \|_\omega$ is equivalent to the standard norm $\| \cdot \|_m$ obtained for $\omega_\mu = 1$,

$$\min_\mu \sqrt{\omega_\mu}\, \|f\|_m \le \|f\|_\omega \le \max_\mu \sqrt{\omega_\mu}\, \|f\|_m \; . \tag{1.2}$$

Thus, the fundamental properties of $H^{m,2}(\mathbb{R})$ extend to $H^{m,2}_\omega(\mathbb{R})$. In particular, $H^{m,2}_\omega(\mathbb{R})$ is continuously embedded into $C^k(\mathbb{R})$ for $0 \le k < m$ (see [1]); i.e., there exist constants $\Gamma(k,m,\omega)$ such that

$$\|\partial^k f\|_\infty \le \Gamma(k,m,\omega)\, \|f\|_\omega \; . \tag{1.3}$$

The *Fourier transform* $\mathcal{F} : f \mapsto \hat{f}$ and its inverse are given by

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \exp(-ixy)\, dx \;, \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(y) \exp(ixy)\, dy \; . \tag{1.4}$$

Strictly speaking, (1.4) is defined for functions $f \in L^1(\mathbb{R})$, but $\mathcal{F}$ can be extended to $L^2(\mathbb{R})$ by a limiting process. Now $\mathcal{F}$ is an *isometry* in $L^2(\mathbb{R})$; i.e.,

$$\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle \;, \quad \|f\| = \|\hat{f}\| \tag{1.5}$$

*Received by the editors January 2, 1996; accepted for publication (in revised form) June 5, 1996. This research was supported by Bundesministerium für Bildung und Forschung (Germany) Projekt 03–HO7STU–2.

http://www.siam.org/journals/sima/28-5/29685.html

[†] Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, 70550 Stuttgart, Germany (reif@mathematik.uni-stuttgart.de).

for all $f, g \in L^2(\mathbb{R})$. Further, if $f(x) \leq C\,(1+|x|)^{-1-\delta}$ and $\hat{f}(y) \leq C\,(1+|y|)^{-1-\delta}$ for some constants C, $\delta > 0$, then the *Poisson summation formula* holds [5, p. 252]:

$$\sum_{j\in\mathbb{Z}} f(j) = \sqrt{2\pi} \sum_{k\in\mathbb{Z}} \hat{f}(2\pi k)\ . \tag{1.6}$$

Denote by $B_{j,n,h}$ the *uniform B-spline* of order $n \in \mathbb{N}$ with knot sequence $h\mathbb{Z}$ and support $\operatorname{supp} B_{j,n,h} = h[j, j+n]$. The *cardinal B-splines* $B_{j,n} = B_{j,n,1}$ are obtained for $h = 1$. Let

$$s(y) := \begin{cases} 2\sin(y/2)/y & \text{for } y \neq 0, \\ 1 & \text{for } y = 0; \end{cases} \tag{1.7}$$

then the Fourier transforms of $B_{j,n}$ and its derivatives are given by

$$\widehat{\partial^\mu B}_{j,n}(y) = (iy)^\mu \exp(-iy(j+n/2))\, s^n(y)/\sqrt{2\pi}; \tag{1.8}$$

see [4, p. 139]. Note that

$$\partial^\mu s^n(y)\big|_{y=2k\pi} = 0 \quad \text{for} \quad k \in \mathbb{Z}\backslash\{0\}\ ,\ 0 \leq \mu < n\ . \tag{1.9}$$

Further, with B_k as the Bernoulli numbers, the Taylor expansion of $1/s$ at the origin is (see [2])

$$1/s(y) = 1 + \sum_{k=1}^{\infty} \frac{(2^{k-1}-1)B_k}{2^{2k}(2k)!}\, y^{2k}\ . \tag{1.10}$$

The *spline space* $S_{n,h} := \operatorname{span} B_{j,n,h}$ consists of linear combinations of B-splines $b := \sum_{j\in\mathbb{Z}} b_j B_{j,n,h}$. With a slight abuse of notation, b will denote both the function and the bi-infinite column vector $[b_j]_{j\in\mathbb{Z}}$ of *B-spline coefficients*.

For an even analytic function f, let $[f]_m := [f_0, \ldots, f_{m-1}]$ be the vector of the first m coefficients of the power series $f(x) = \sum_{j=0}^\infty f_j x^{2j}$. Denote by $*$ the *convolution operator* and by a^{*k} the $(k-1)$-fold convolution of a vector a with itself; i.e.,

$$* : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}^m\ , \quad (a*b)_\nu := \sum_{\mu=0}^{\nu} a_{\nu-\mu} b_\mu, \tag{1.11}$$

$$a^{*k} := a * a^{*(k-1)}\ , \quad a^{*1} := a; \tag{1.12}$$

then

$$[fg]_m = [f]_m * [g]_m\ , \quad [f^k]_m = [f]_m^{*k}\ . \tag{1.13}$$

2. Weights providing orthonormality. The cardinal B-splines $B_{j,n}$ have compact support and a piecewise constant derivative of order $n-1$; thus $B_{j,n} \in H_\omega^{n-1,2}(\mathbb{R})$ for arbitrary ω. It is the main result of this paper that $\omega = \omega(n)$ can be chosen such that $\{B_{j,n},\ j \in \mathbb{Z}\}$ becomes an orthonormal sequence.

THEOREM 2.1. *The sequence $\{B_{j,n}, j \in \mathbb{Z}\}$ is orthonormal in $H_{\omega(n)}^{n-1,2}(\mathbb{R})$ if and only if*

$$\omega(n) := [1/s]_n^{*2n}\ . \tag{2.1}$$

In particular, $\omega_\mu(n) > 0$ *for all* $n \in \mathbb{N}$ *and* $0 \le \mu < n$.

Proof. Since $(B_{j_1,n}, B_{j_2,n})_\omega = (B_{j_2,n}, B_{j_1,n})_\omega = (B_{0,n}, B_{|j_2-j_1|,n})_\omega$, it suffices to consider inner products of type $(B_{0,n}, B_{j,n})_\omega, j \ge 0$. For $j \ge n$ the supports of $B_{0,n}$ and $B_{j,n}$ are disjoint; hence $(B_{0,n}, B_{j,n})_\omega = 0$. For $j = 0, \dots, n-1$, we obtain using (1.5) and (1.8)

$$\begin{aligned}(B_{0,n}, B_{j,n})_\omega &= \sum_{\mu=0}^{n-1} \omega_\mu \, \langle \partial^\mu B_{0,n}, \partial^\mu B_{j,n} \rangle \\ &= \sum_{\mu=0}^{n-1} \frac{\omega_\mu}{2\pi} \int_{-\infty}^{\infty} y^{2\mu} s^{2n}(y) \exp(ijy) \, dy \\ &= \sum_{\mu=0}^{n-1} (-1)^\mu \, \omega_\mu \, \partial^{2\mu} B_{0,2n}(j+n) \, . \end{aligned} \tag{2.2}$$

Define the column vector e by $e_j := \delta_{j,0}$ and the $n \times n$-matrices P, Q by

$$P_{j,\mu} := (-1)^\mu \partial^{2\mu} B_{0,2n}(j+n), \tag{2.3}$$

$$Q_{\nu,j} := \frac{2(-1)^\nu j^{2\nu}}{(2\nu)!} - \delta_{\nu,0}\delta_{j,0} \, . \tag{2.4}$$

Scaling the rows of Q appropriately yields the Vandermonde-matrix with entries $j^{2\nu}$. Thus Q is invertible, and orthonormality of $\{B_{j,n}, \, j \in \mathbb{Z}\}$ is equivalent to

$$QP\omega = Qe = e \, . \tag{2.5}$$

For computing the product matrix $R := QP$ the summation array $j = 0, \dots, n-1$ can be transformed to $\mathbb{Z}$ exploiting $B_{0,2n}(j+n) = B_{0,2n}(-j+n)$ and $\operatorname{supp} B_{0,2n} = [0, 2n]$,

$$R_{\nu,\mu} = \sum_{j=0}^{n-1} Q_{\nu,j} P_{j,\mu} = \frac{(-1)^{\nu+\mu}}{(2\nu)!} \sum_{j \in \mathbb{Z}} j^{2\nu} \partial^{2\mu} B_{0,2n}(j+n) \, . \tag{2.6}$$

So, (1.6) becomes applicable, and we obtain the following using (1.9):

$$\begin{aligned} R_{\nu,\mu} &= \frac{1}{(2\nu)!} \sum_{k \in \mathbb{Z}} \partial^{2\nu} \big(y^{2\mu} s^{2n}(y) \big)\big|_{y=2k\pi} \\ &= \frac{1}{(2\nu)!} \sum_{k \in \mathbb{Z}} \sum_{\ell=0}^{2\nu} \binom{2\nu}{\ell} \partial^\ell \big(y^{2\mu} \big)\big|_{y=2k\pi} \partial^{2\nu-\ell} \big(s^{2n}(y) \big)\big|_{y=2k\pi} \\ &= \frac{1}{(2\nu)!} \sum_{\ell=0}^{2\nu} \binom{2\nu}{\ell} \partial^\ell \big(y^{2\mu} \big)\big|_{y=0} \partial^{2\nu-\ell} \big(s^{2n}(y) \big)\big|_{y=0} \\ &= \begin{cases} 0 & \text{for } \nu < \mu, \\ \frac{1}{(2\nu-2\mu)!} \, \partial^{2(\nu-\mu)} \big(s^{2n}(y) \big)\big|_{y=0} & \text{for } \nu \ge \mu \, . \end{cases} \end{aligned} \tag{2.7}$$

Consequently, $R\omega = [s^{2n}]_n * \omega$, and setting $\omega(n) = [1/s]_n^{*2n}$ yields

$$R\omega(n) = [s^{2n}]_n * [1/s]_n^{*2n} = [s^{2n}]_n * [1/s^{2n}]_n = [1]_n = e \, . \tag{2.8}$$

The solution is positive for all $n \in \mathbb{N}$ by (1.10) and unique since $\det R = 1$. □

TABLE 2.1
Weights $\omega(n)$ for $n \leq 7$.

n	$\omega_0(n)$	$\omega_1(n)$	$\omega_2(n)$	$\omega_3(n)$	$\omega_4(n)$	$\omega_5(n)$	$\omega_6(n)$
1	1						
2	1	$\frac{1}{6}$					
3	1	$\frac{1}{4}$	$\frac{1}{30}$				
4	1	$\frac{1}{3}$	$\frac{7}{120}$	$\frac{1}{140}$			
5	1	$\frac{5}{12}$	$\frac{13}{144}$	$\frac{41}{3024}$	$\frac{1}{630}$		
6	1	$\frac{1}{2}$	$\frac{31}{240}$	$\frac{139}{6048}$	$\frac{479}{151200}$	$\frac{1}{2772}$	
7	1	$\frac{7}{12}$	$\frac{7}{40}$	$\frac{311}{8640}$	$\frac{37}{6480}$	$\frac{59}{79200}$	$\frac{1}{12012}$

Table 2.1 shows the weights $\omega(n)$ for $n \leq 7$. The result of Theorem 2.1 can be readily generalized to uniform B-splines by scaling.

COROLLARY 2.2. *The sequence $\{B_{j,n,h}, j \in \mathbb{Z}\}$ is orthonormal in $H^{n-1,2}_{\omega(n,h)}(\mathbb{R})$ if and only if*

$$\omega_\mu(n,h) := h^{2\mu-1}\omega_\mu(n)\ . \tag{2.9}$$

3. An application.

A typical application, where the orthonormality of B-splines is of advantage, is the approximation of functions by splines.

THEOREM 3.1. *For $f \in H^{n-1,2}(\mathbb{R})$, consider the approximation problem*

$$\|f-g\|_{\omega(n,h)} \to \min\ , \quad g \in S_{n,h}\ . \tag{3.1}$$

The B-spline coefficients of the solution $Q_h f = \sum_{j\in\mathbb{Z}}(Q_h f)_j B_{j,n,h}$ are given by

$$(Q_h f)_j := (f, B_{j,n,h})_{\omega(n,h)}\ . \tag{3.2}$$

Thus, the projection $Q_h := H^{n-1,2}(\mathbb{R}) \to S_{n,h}$ is local in the sense that $(Q_h f)(x)$ depends only on the restriction of f to the interval $[x-nh, x+nh]$.

Proof. The proof is trivial. □

The approximation error satisfies the following estimates.

THEOREM 3.2. *For $f \in H^{n-1,2}(\mathbb{R}) \cap H^{n,\infty}(\mathbb{R})$ and $k \in [0,\ldots,n-1]$ there exists a constant C depending only on n and k such that*

$$\|\partial^k(f-Q_h f)\|_\infty \leq C\, h^{n-k}\, \|\partial^n f\|_\infty\ . \tag{3.3}$$

Moreover, for $f \in H^{n-1,2}(\mathbb{R})$ and $0 \leq k \leq n-2$,

$$\begin{aligned} \|\partial^k(f-Q_h f)\|_\infty &\leq \Gamma(k,n-1,\omega(n,h))\|f-Q_h f\|_{\omega(n,h)} \\ &= \Gamma(k,n-1,\omega(n,h))\left(\|f\|^2_{\omega(n,h)} - \sum_{j\in\mathbb{Z}} f_j^2\right)^{1/2} \end{aligned} \tag{3.4}$$

with the constant declared in (1.3).

Proof. Let $P_h := C^n(\mathbb{R}) \mapsto S_{n,h}$ be a standard quasi interpolant of order n. Set $\Delta_h := f - P_h f$; then

$$\|\partial^k \Delta_h\|_\infty \leq C_1\, h^{n-k}\|\partial^n f\|_\infty \tag{3.5}$$

with C_1 some constant depending only on n and k (see [4, p. 229]). $S_{n,h}$ is invariant under Q_h (i.e., $Q_h P_h = P_h$), so

$$\begin{aligned} \|\partial^k(f - Q_h f)\|_\infty &= \|\partial^k(\Delta_h - Q_h\Delta_h)\|_\infty \\ &\le C_1\, h^{n-k}\|\partial^n f\|_\infty + \|\partial^k(Q_h\Delta_h)\|_\infty\ . \end{aligned} \tag{3.6}$$

Using $\int_{\mathbb{R}} \partial^\nu B_{j,n,h}(x)\,dx = h^{1-\nu}\int_{\mathbb{R}} \partial^\nu B_{0,n}(x)\,dx$, we obtain for the second summand

$$\begin{aligned} \|\partial^k Q_h\Delta_h\|_\infty &\le \sum_{j\in\mathbb{Z}} |(\Delta_h, B_{j,n,h})_{\omega(n,h)}|\;\|\partial^k B_{j,n,h}\|_\infty \\ &\le h^{-k}\,\|\partial^k B_{0,n}\|_\infty \sup_{j\in\mathbb{Z}} |(\Delta_h, B_{j,n,h})_{\omega(n,h)}| \\ &\le \|\partial^k B_{0,n}\|_\infty \sum_{\nu=0}^{n-1} h^{2\nu-k-1}\,\omega_\nu(n) \sup_{j\in\mathbb{Z}} |\langle\partial^\nu\Delta_h, \partial^\nu B_{j,n,h}\rangle| \\ &\le C_2\, h^{n-k}\,\|\partial^n f\|_\infty\ . \end{aligned} \tag{3.7}$$

The constant C_2 also depends only on n and k; thus, (3.3) holds with $C := C_1 + C_2$. (3.4) is an immediate consequence of (1.3). □

Since $Q_h f$ depends only locally on f, the domain of Q_h can be extended significantly.

Corollary 3.3. *For $f \in H^{n-1,2}_{\text{loc}}(\mathbb{R}) \cap H^{n,\infty}_{\text{loc}}(\mathbb{R})$ (i.e., $f_{|r} \in H^{n-1,2}(r) \cap H^{n,\infty}(r)$ for any compact interval $r \subset \mathbb{R}$), the operator Q_h is well defined by* (3.2)*, and in analogy to* (3.3)*, the estimate*

$$\|\partial^k(f - Q_h f)(x)\| \le C\,h^{n-k}\,\|\partial^n f_{|[x-nh,x+nh]}\|_\infty \tag{3.8}$$

holds for $0 \le k < n$.

The numerical evaluation of (3.2) can be made efficient by using the identity

$$\langle\partial^\mu f, \partial^\mu B_{j,n,h}\rangle = \begin{cases} (-1)^\mu\,\langle f, \partial^{2\mu} B_{j,n,h}\rangle & \text{if } 2\mu < n, \\ (-1)^\mu h n!\,[hj,\dots,h(j+n)]\partial^{2\mu-n} f & \text{if } 2\mu \ge n. \end{cases} \tag{3.9}$$

The benefit is that the derivatives of f have to be evaluated solely at the knots and not at the multitude of arguments required by the quadrature scheme. (3.9) is based on repeated integration by parts. In the second case, the process stops, when the B-spline has been transformed to $\partial^{n-1}B_{j,n,h}$, which is piecewise constant. Thus, the integral can be evaluated and yields the given divided difference. A more detailed proof is not very instructive in this context.

In many applications, approximation is subject to a finite number of linear constraints, say Hermite interpolation at certain points. It turns out that the solution of such a problem is simply obtained by an orthogonal projection of the unconstrained approximant $Q_h f$ on the feasible set.

Theorem 3.4. *For $f \in H^{n-1,2}(\mathbb{R})$, consider the constrained approximation problem*

$$\|f - g\|_{\omega(n,h)} \to \min\ , \quad \Lambda g = \lambda\ , \quad g \in S_{n,h}\ , \tag{3.10}$$

where Λ is a full rank matrix with k absolutely summable bi-infinite rows. The solution $Q_h^\Lambda f = \sum_{j\in\mathbb{Z}} (Q_h^\Lambda f)_j B_{j,n,h}$ is given by

$$(Q_h^\Lambda f) := (Q_h f) - \Lambda^T(\Lambda\Lambda^T)^{-1}(\Lambda\, Q_h f - \lambda)\ . \tag{3.11}$$

Proof. Note that $\Lambda\, Q_h f$ is well defined since $(Q_h f)_j$ is bounded by

$$|(Q_h f)_j| \leq \|B_{j,n,h}\|_{\omega(n,h)} \|f\|_{\omega(n,h)} = \|B_{0,n,h}\|_{\omega(n,h)} \|f\|_{\omega(n,h)} \,. \tag{3.12}$$

Introducing the vector p of Lagrange multipliers, (3.10) is equivalent to

$$Q_h^\Lambda f + \Lambda^T p = Q_h f, \tag{3.13}$$

$$\Lambda\, Q_h^\Lambda f = \lambda \,. \tag{3.14}$$

Multiplication of the first equation by Λ and substitution yields $(\Lambda\Lambda^T)p = \Lambda\, Q_h f - \lambda$. The $k \times k$-matrix $(\Lambda\Lambda^T)$ is invertible; thus (3.11) follows. □

The benefits of approximation in $H^{m,2}_{\omega(n,h)}(\mathbb{R})$ with orthonormal B-splines are evident. First, Q_h is local and can be computed explicitly and efficiently using (3.9). Second, (3.3) indicates that the approximation order of Q_h is optimal. This property is shared by a large family of quasi interpolants, but Q_h stands out due to the fact that it is the best approximation with respect to a reasonable inner product, measuring the total deviation of function values and certain derivatives. Third, (3.4) guarantees that if the approximation is good in $H^{n-1,2}_{\omega(n,h)}(\mathbb{R})$, then the maximum norm of the error is small. This feature is of particular importance in many applications, where the absolute error is required to be smaller than a given tolerance, everywhere. Fourth, constrained approximation simply splits into solving the unconstrained problem and a subsequent projection step.

Acknowledgment. The author thanks Klaus Höllig for helpful comments.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure and Applied Mathematics Series, Academic Press, 1978.

[2] I. N. Bronstein and K. A. Semendjajew, *Taschenbuch der Mathematik*, 19th ed., Verlag Harri Deutsch, Leipzig, 1979.

[3] C. de Boor, *A Practical Guide to Splines*, Applied Mathematical Sciences, Springer-Verlag, Berlin, 1978.

[4] L. L. Schumaker, *Spline functions: Basic theory*, in Pure and Applied Mathematics, Wiley-Interscience, New York, 1981.

[5] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1975.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1265–1293, November 1997

GINZBURG–LANDAU VORTICES: DYNAMICS, PINNING, AND HYSTERESIS [*]

FANG-HUA LIN[†] AND QIANG DU[‡]

Abstract. In this paper, we consider three problems related to the mathematical study of vortex phenomena in superconductivity based on the G–L models. First, we study the long-time behavior of the solutions of the time-dependent Ginzburg–Landau equations. Then we describe results concerning the pinning effect of thin regions in a variable thickness thin film. Finally, we prove the existence of vortex-like solutions to the steady state Ginzburg–Landau equations and study the hysteresis phenomenon near the lower critical field.

Key words. superconductivity, Ginzburg–Landau equations, vortex solution, long-time behavior, vortex pinning, hysteresis

AMS subject classifications. 35B40, 35A20, 35J65, 35K99, 82D55

PII. S0036141096298060

1. Introduction.

Below the critical temperature T_c, the response of a superconducting material to an externally imposed magnetic field is most conveniently described by the diagram given in Figure 1, which shows the minimum energy state of the superconductor as a function of H_0, the applied magnetic field, and the dimensionless material parameter κ (known as the Ginzburg–Landau parameter). The parameter κ determines the type of superconducting material [10]. For $\kappa < \frac{1}{\sqrt{2}}$, type-I superconductors, there is a critical magnetic field H_C below which the material will be the superconducting Meissner state but above which it will be in the normal state. For $\kappa > \frac{1}{\sqrt{2}}$, type-II superconductors, there is a third state known as the mixed (or vortex) state. The vortex state consists of many normal filaments embedded in a superconducting matrix. Each of these filaments carries with it a quantized amount of magnetic flux and is circled by a vortex of superconducting current. Thus, these filaments are often know as vortex lines. One of the most challenging problems to mathematicians working on the superconductivity models is to understand vortex phenomena in type-II superconductors, which include the recently discovered high-temperature superconductors.

The transition from the normal state to the vortex state takes place by a bifurcation as the magnetic field is lowered through some critical value H_{C_2}. The critical field H_{C_1}, on the other hand, is calculated so that at this field the energy of the wholly superconducting solution becomes equal to the energy of the single vortex filament solution for infinite superconductors.

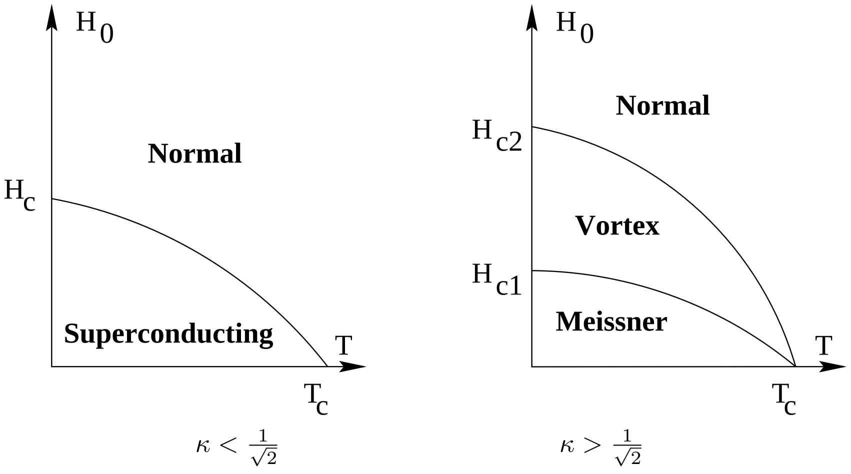
The vortex structures have been studied extensively on the mezoscale by using the well-known Ginzburg–Landau (G–L) models of superconductivity [10, 13, 14]. The existence of vortex-like solutions for the full nonlinear G–L equations has been investigated by researchers using methods ranging from asymptotic analysis to numerical

[*] Received by the editors January 31, 1996; accepted for publication (in revised form) August 28, 1996.

http://www.siam.org/journals/sima/28-6/29806.html

[†] Courant Institute, New York University, New York, NY 10012 (linf@cims.nyu.edu). The research of this author was partially supported by NSF grant MS-9401546.

[‡] Department of Mathematics, Iowa State University, Ames, IA 50011-2064. Current address: Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (qdu@iastate.edu). The research of this author was partially supported by NSF grant MS-9500718.



$\kappa < \frac{1}{\sqrt{2}}$ $\kappa > \frac{1}{\sqrt{2}}$

FIG. 1. *The various states of superconductors.*

simulations; however, it has not been justified by rigorous mathematical analysis. Much progress has been made in recent years [3, 20, 21] to establish a mathematical framework for a rigorous description of both the static and dynamic properties of the vortex solutions; in particular, as the coherence length tends to zero (κ goes to infinity), various results have been obtained. From a technological point of view, this is of interest since recently discovered high critical temperature superconductors are known to have large values of κ, say κ in excess of 50.

Vortex lines may move as a result of internal interactions between these filaments and external forces (due to applied fields or thermal fluctuations) acting on them. Unfortunately, such vortex motion in an applied magnetic field induces an effective electrical resistance in the material and, thus, a loss of superconductivity. Therefore, it is crucial to understand the dynamic of these vortex lines. At the same time, one is interested in studying mechanisms that can pin the vortices at a fixed location, i.e., prevent their motion. Various such mechanisms have been advanced by physicists, engineers, and material scientists. For example, normal (nonsuperconducting) impurities in an otherwise superconducting material sample are believed to provide sites at which vortices are pinned. Likewise, regions of the sample that are thin relative to other regions are also believed to provide pinning sites. These mechanisms have been introduced into the general G–L framework to derive various variants of the original G–L models of superconductivity. Numerical simulation clearly suggests the pinning effect.

In this paper, we deal with three independent problems, yet all of them are related to the study of vortex phenomena. First we study the long-time behavior of the solutions of the time-dependent G–L equations and the main result is Theorem 2.1. Then we describe results (Theorems 3.1–3.5) concerning the pinning effect of thin regions in a variable thickness thin film, based on the models developed in [6, 12], and finally, we prove the existence of vortex-like solutions to the steady state G–L equations (Lemma 4.1) and study the hysteresis phenomenon near the lower critical field.

Before addressing the above problems, we introduce the notation and the models that will be used in the paper. The starting point of our study is the phenomenological model due to Ginzburg and Landau for superconductivity in isotropic, homogeneous material samples. Let Ω be a smooth bounded domain in $\mathbb{R}^3$, occupied by the superconducting material. By ignoring the effect of the region exterior to the sample, the steady state model can be stated as a minimization problem of the free energy functional

$$\begin{aligned}\mathcal{G}(\psi, \mathbf{A}) = \int_\Omega \Bigg\{ & f_n + a|\psi|^2 + \frac{b}{2}|\psi|^4 + \frac{1}{2m_s}\left|\left(i\bar{h}\nabla + \frac{e_s}{c}\mathbf{A}\right)\psi\right|^2 \\ & + \frac{\mu_s}{8\pi}\mathbf{h}\cdot(\mathbf{h} - 2\mathbf{H}_0)\Bigg\}\, d\Omega\,, \end{aligned} \tag{1}$$

where f_n is the free energy density of the nonsuperconducting state in the absence of a magnetic field, ψ is the (complex-valued) superconducting order parameter, $\mathbf{A}$ is the magnetic vector potential, $\mathbf{h} = (1/\mu_s)\text{curl}\mathbf{A}$ is the magnetic field, $\mathbf{H}_0$ is the applied magnetic field, a and b are constants whose values depend on the temperature and such that $b > 0$, e_s is the mass of the superconducting charge carriers which is twice the electronic charge e, c is the speed of light, μ_s is the permeability, and $2\pi\bar{h}$ is Planck's constant. It can be rewritten in nondimensionalized form:

$$\mathcal{G}(\psi, \mathbf{A}) = \int_\Omega \left(\frac{1}{2}(1 - |\psi|^2)^2 + \left|\left(\frac{i}{\kappa}\nabla + \mathbf{A}\right)\psi\right|^2 + |\text{curl}\,\mathbf{A} - \mathbf{H}_0|^2\right) d\mathbf{x}, \tag{2}$$

where κ is the so-called G–L parameter.

The functional $\mathcal{G}(\psi, \mathbf{A})$ has an interesting gauge invariance property and the minimization of $\mathcal{G}$ in appropriate functional spaces gives the following system of nonlinear differential equations that are named the G–L equations:

$$\left(\frac{i}{\kappa}\nabla + \mathbf{A}\right)^2 \psi - \psi + |\psi|^2\psi = 0 \quad \text{in } \Omega, \tag{3}$$

$$\text{curl}\,\text{curl}\,\mathbf{A} = \frac{i}{2\kappa}(\psi\nabla\psi^* - \psi^*\nabla\psi) - |\psi|^2\mathbf{A} \quad \text{in } \Omega, \tag{4}$$

along with natural boundary conditions

$$\text{curl}\,\mathbf{A} \wedge \mathbf{n} = \mathbf{H}_0 \wedge \mathbf{n} \quad \text{on } \partial\Omega \tag{5}$$

and

$$\left(\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi\right)\cdot\mathbf{n} = 0 \quad \text{on } \partial\Omega, \tag{6}$$

where $\mathbf{n}$ is the exterior normal to the boundary $\partial\Omega$. The G–L vortices are represented by the zeros of the complex order parameter ψ. In section 4, we will prove the existence of vortex solutions to the above system when κ is large and the applied field is near the lower critical field.

Equations (3)–(4) are the steady state G–L equations. The time-dependent G–L model is often described by the Gorkov–Eliashberg evolution equation [17]:

$$\left\{\begin{aligned} & \eta\frac{\partial\psi}{\partial t} + i\,\eta\,\kappa\,\Phi\,\psi + \left(\frac{i}{\kappa}\,\nabla + \mathbf{A}\right)^2 \psi - \psi + |\psi|^2\,\psi = 0, \\ & \frac{\partial\mathbf{A}}{\partial t} + \nabla\Phi + \text{ curl curl } \mathbf{A} = -\frac{i}{2\kappa}\left(\psi^*\,\nabla\psi - \psi\,\nabla\psi^*\right) - \mathbf{A}|\psi|^2. \end{aligned}\right. \tag{7}$$

Here, Φ denotes the (real) scalar electric potential, η is a relaxation parameter, and ψ^* denotes the complex conjugate of ψ. For simplicity, we take $\eta = 1$ in the rest of the paper.

The system is supplemented by the initial and boundary conditions

$$\psi\,(x,0) = \psi_0(x), \quad \mathbf{A}(x,0) = \mathbf{A}_0(x), \qquad x \in \Omega; \tag{8}$$

$$\begin{cases} \left(\dfrac{i}{\kappa}\nabla + \mathbf{A}\right)\psi \cdot \mathbf{n} = 0, \\ \qquad \text{curl}\,\mathbf{A}\ \wedge\ \mathbf{n} = \mathbf{H}_0 \wedge \mathbf{n}, \\ \left(\dfrac{\partial \mathbf{A}}{\partial t} + \nabla\Phi\right)\cdot\ \mathbf{n} = \vec{E}\cdot\mathbf{n}\ =\ 0 \quad \text{on } \partial\Omega. \end{cases} \tag{9}$$

Note that (7) and (9) are gauge invariant [11], in the sense that if $(\psi,\ \mathbf{A},\ \Phi)$ is a solution, then so is $(\psi_\chi,\ \mathbf{A}_\chi,\ \Phi_\chi)$, where

$$\psi_\chi = \psi\, e^{i\,\kappa\,\chi}, \quad \mathbf{A}_\chi = \mathbf{A} + \nabla\chi, \quad \Phi_\chi = \Phi - \frac{\partial\chi}{\partial t}.$$

The dynamics of vortices can be determined from the solutions of the time-dependent equations (7)–(9). The long-time asymptotic behavior of solutions of equations (7)–(9) as $t \to \infty$ will be studied later in section 2.

For type-II superconductors, the minimizers of $\mathcal{G}$ are believed to exhibit vortex structures. Numerical experiments show that for large values of κ and moderate field strength, the number of vortices could be exceedingly large even for a small sample size in actual physical scale. Thus, resolving the vortex phenomenon by using the full G–L equations remains computationally intensive.

Various simplifications have been made to reduce the complexity. For thin films of superconducting material, a two-dimensional model has been developed [6, 12] that can account for thickness variations through an averaging process. The model is given by the following minimization problem:

$$\mathcal{G}^a_\epsilon(\psi) = \int_\Omega a(\mathbf{x})\left(|(\nabla - i\mathbf{A}_0)\psi|^2 + \frac{1}{2\epsilon^2}(1-|\psi|^2)^2\right)d\mathbf{x}, \tag{10}$$

where Ω denotes the platform of the film, $a(\mathbf{x})$ measures the relative thickness of the film, and $\mathbf{A}_0$ is a prescribed vector potential due to the normal (to film) component of the applied field. The role of the Ginzburg–Landau parameter κ is assumed by the parameter $\epsilon(\propto 1/\kappa)$.

It was proved that for fixed ϵ, the minimizers of the above problem, along with the prescribed vector potential $\mathbf{A}_0$, provide the leading order approximation to the solution of the three-dimensional problem [6]. The creation and interaction of vortices based on the above model is connected to the prescribed magnetic potential. The number of vortices cannot be prescribed a priori, independently of $\mathbf{A}_0$. To simplify the analysis further, a simpler problem, in which the number of vortices is prescribed and the magnetic potential is ignored, can be studied. By rescaling the spatial variables, one may consider the minimization of the functional

$$\mathcal{F}^a_\epsilon(\psi) = \int_\Omega a(\mathbf{x})\left(|\nabla\psi|^2 + \frac{1}{2\epsilon^2}(1-|\psi|^2)^2\right)d\mathbf{x} \tag{11}$$

with the boundary condition

$$\psi(\mathbf{x}) = g(\mathbf{x}) \quad \text{for } \mathbf{x} \in \partial\Omega, \tag{12}$$

where g is smooth with $|g(\mathbf{x})| = 1, \mathbf{x} \in \partial\Omega$. This can be viewed as a generalization of the problem studied in [3, 20] in which $a(\mathbf{x}) \equiv 1$, i.e.,

$$\mathcal{F}_\epsilon(\psi) = \int_\Omega \left(|\nabla\psi|^2 + \frac{1}{2\epsilon^2}(1 - |\psi|^2)^2 \right) d\mathbf{x}. \tag{13}$$

The pinning effect of the variable thickness will be studied in section 3 by examining the properties of minimizers of (10) and (11).

The rest of the paper is devoted to the problems we have discussed above.

2. The uniqueness of the asymptotic limit. The global existence and uniqueness (up to gauge transformations) of classical solutions of (7)–(9) have been studied by various authors, e.g., [7, 11, 28]. The dynamics of vortices (including the simpler case which ignores the magnetic field) have been studied by [15, 16, 23, 24, 25] and recently proved in [22]. In [28], the long-time behavior, in particular the existence of the global attractor, is also investigated. Here, we shall sketch the proof of the asymptotic stability result, which shows that, as $t \to \infty$, (7)–(9) has a unique asymptotic limit up to gauge transformations.

As in [11], one may choose the so-called zero electric potential gauge for system (7)–(9). To do so, one must solve

$$\frac{\partial\chi}{\partial t} = \Phi \tag{14}$$

and, at $t = 0$,

$$\Delta\chi = -\operatorname{div}\mathbf{A} \ \text{ in } \Omega \quad \text{with} \quad \nabla\chi \cdot \mathbf{n} = -\mathbf{A}\cdot\mathbf{n} \ \text{ on } \partial\Omega. \tag{15}$$

Thus, in this gauge, $\Phi \equiv 0$ and system (1) reduces to the gradient flow of the energy functional:

$$\begin{aligned} E(\psi, \mathbf{A}) = \tfrac{1}{2}\int_\Omega \Bigg[\left| \left(\frac{i}{\kappa}\nabla + \mathbf{A} \right)\psi \right|^2 + \tfrac{1}{2}\,(|\psi|^2 - 1)^2 \\ + |\operatorname{curl}\mathbf{A} - \mathbf{H}_0|^2 \Bigg] dx. \end{aligned} \tag{16}$$

The initial condition satisfies $\operatorname{div}\mathbf{A}(0) = 0$ and $\mathbf{A}(0)\cdot\mathbf{n} = 0$. Moreover, we make a physically meaningful assumption that $\|\psi(0)\|_\infty \le 1$. Let $v = (\psi, \mathbf{A})$. As shown in [11] and [28], the flow

$$\frac{dv}{dt} = -\ \text{grad}\ \ E(v), \qquad v(0) = v_0, \tag{17}$$

has a global classical solution. Let $\mathbf{V} = \mathcal{H}^2(\Omega) \times \mathbf{H}^2(\Omega)$, where $\mathcal{H}^2(\Omega)$ and $\mathbf{H}^2(\Omega)$ are spaces of functions whose components (or real and imaginary parts) are in the standard Sobolev space $H^2(\Omega)$.

Our main result concerning (17) now follows.

Theorem 2.1.

$$V_\infty = \lim_{t\to\infty} v(t) \qquad \textit{exists in } \mathbf{V}\ .$$

To describe the idea, we start with the ODE

$$\begin{cases} \dfrac{dx}{dt} &= -\text{ grad } f(x), \qquad x \in \mathbb{R}^N \\ x(0) &= x_0 \, . \end{cases} \tag{18}$$

We assume $f \in C^2(\mathbb{R}^N)$, $\nabla \mathrm{f}(0) = 0$, and x_0 is close to 0.

Case (i). If $A = \nabla^2 f(0)$ is positive definite, then $x(t) \to 0$ (at an exponential rate) as $t \to +\infty$.

This result is standard. One can calculate

$$\frac{d}{dt} |\dot{x}|^2 = -2\langle \nabla^2 f(x) \cdot \dot{x}, \ \dot{x} \rangle \leq -2\lambda \, |\dot{x}|^2 \, .$$

Here we shall assume $|x|(t) \leq \delta_0$, and $(\nabla^2 f(x)) \geq \lambda I$ whenever $|x| \leq \delta_0$. Thus $|\dot{x}(t)| \leq |\dot{x}(0)| \, e^{-\lambda t} \quad \forall t \geq 0$ and $|x(t)| \leq |x_0| + \frac{1}{\lambda}|\dot{x}(0)| = |x_0| + \frac{1}{\lambda}|\nabla f(x_0)|$. We shall always assume x_0 is so close to the origin that $|x_0| + \frac{1}{\lambda}|\nabla f(x_0)| < \delta_0$. Then the assumption $|x(t)| \leq \delta_0$ is true for all $t > 0$ and, thus, $x(t) \to 0$ at an exponential rate as $t \to +\infty$.

Case (ii). det $(A) \neq 0$. Then one has that

$$-\frac{d}{dt} \left(f(x) - f(0)\right)^{1/2} = \frac{1}{2} \, \frac{|\nabla f(x)| \ |\dot{x}|}{(f(x) - f(0))^{1/2}}, \quad \text{if } f(x) > f(0), \tag{19}$$

$$\frac{d}{dt} \left(f(0) - f(x)\right)^{1/2} = \frac{1}{2} \, \frac{|\nabla f(x)| \ |\dot{x}|}{|f(x) - f(0)|^{1/2}}, \quad \text{if } f(x) < f(0). \tag{20}$$

We obtain the following: *either* there is a $T \in (0, \infty)$ such that

$$f(x)(T) \leq f(0) - \delta_0 \qquad (\text{for some } \ \delta_0 > 0)$$

or

$$\lim_{t \to +\infty} \ x(t) = x_\infty \quad \text{ exists.}$$

Indeed, if x is close to 0, then

$$\frac{|\nabla f(x)|}{|f(x) - f(0)|^{1/2}} \ \geq \ \frac{\lambda_{\min}}{\sqrt{2\lambda_{\max}}} \ = \ C(A) > 0.$$

Here $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum, respectively, eigenvalues of $(A^T A)^{1/2}$. Therefore,

$$\int_0^\infty |\dot{x}(t)| \, dt \ \leq \ \frac{2\delta_0^{\,1/2}}{C(A)}$$

by (19)–(20) whenever $f(x)(t) \geq f(0) - \delta_0$ for all $t \geq 0$. In such a case the conclusion $\lim_{t \to +\infty} x(t) = x_\infty$ follows (in fact $x_\infty = 0$) as for Case (i).

Case (iii). det $(A) = 0$ and $f(x)$ is real analytic in B_1. We have the following well-known estimate (cf. [26, 27]). There are two positive constants $\theta_0, \sigma_0 \in (0,1)$ depending on f such that

$$|f(x) - f(0)|^{\theta_0} \ \leq \ |\text{ grad } \ f(x)| \quad \text{whenever } x \in B_{\sigma_0}(0), \tag{21}$$

$\nabla f(0) = 0$. Then, as for Case (ii), one has that *either* there is a $T \in (0, \infty)$ such that

$$f(T) \le f(0) - \delta_0$$

or

$$\lim_{t \to \infty} x(t) = x_\infty \quad \text{exists.}$$

In [27], Simon considered the case

$$E(u) = \int_M F(x, u, \nabla u)\, dx, \tag{22}$$

where M is a compact manifold without boundary and

$$\begin{cases} \dot{u} & = - \text{ grad } \ E(u) \ \equiv \ \mathcal{M}(u), \\ u(0) & = \ u_0 \simeq 0, \qquad \mathcal{M}(0) = 0. \end{cases} \tag{23}$$

Here, F is assumed to be analytic in both u and ∇u for u, ∇u near $\underline{0}$.

Suppose L is elliptic, $Lv = \frac{d}{ds}|_{s=0}\, \mathcal{M}(sv)$. Then *either* there is a $T \in (0, \infty)$ such that

$$E(u(T)) \le E\,(0) - \delta_0 \quad \text{for some } \delta_0 > 0$$

or

$$u_\infty = \lim_{t \to \infty} u(t) \quad \text{exists.}$$

To apply the above idea to the time-dependent G–L equations, we need to use the following estimate given in [11] (also see [28] for similar results).

LEMMA 2.2. *Let $v = (\psi, \mathbf{A})$ be the solution of the time-dependent G–L equations* (17). *Then*

$$\int_t^T \Big[(\dot{\psi}(s), \dot{\psi}(s)) + (\dot{\mathbf{A}}(s), \dot{\mathbf{A}}(s)) \Big]\, ds + E(\psi(T), \mathbf{A}(T)) = E(\psi(t), \mathbf{A}(t)) \tag{24}$$

for any $T > t > 0$.

The following lemma will enable us to apply the above conclusion of Simon.

LEMMA 2.3. *Let $(\psi, \mathbf{A})$ satisfy*

$$\operatorname{curl} \mathbf{A} \wedge \mathbf{n} = \mathbf{H}_0 \wedge \mathbf{n}\,, \quad \mathbf{A} \cdot \mathbf{n} = 0, \quad \textit{and} \quad \nabla\psi \cdot \mathbf{n} = 0 \quad \textit{on } \partial\Omega.$$

Let $E(\psi, \mathbf{A})$ be defined by (16) *and $(\psi_*, \mathbf{A}_*)$ be a steady state solution of the G–L equations in the gauge $\operatorname{div} \mathbf{A}_* = 0$ in Ω and $\mathbf{A}_* \cdot \mathbf{n} = 0$ on $\partial\Omega$. Then there exist constants $\theta_0, \sigma_0 \in (0, 1)$ such that*

$$|E(\psi, \mathbf{A}) - E(\psi_*, \mathbf{A}_*)|^{\theta_0} \le \| \text{ grad } \ E(\psi, \mathbf{A})\|_{L^2(\Omega)} \tag{25}$$

for any $\|(\psi, \mathbf{A}) - (\psi_*, \mathbf{A}_*)\|_{C^2(\Omega)} \le \sigma_0$.

Proof. Let

$$G\,(\psi,\, \mathbf{A}) = E\,(\psi,\, \mathbf{A}) + \tfrac{1}{2} \int_\Omega \big[|\operatorname{div} \mathbf{A}|^2 \,\big] \, dx. \tag{26}$$

By assumption, since $(\psi_*, \mathbf{A}_*)$ is a steady state solution of the G–L equations in the gauge $\operatorname{div} \mathbf{A}_* = 0$ in Ω and $\mathbf{A}_* \cdot \mathbf{n} = 0$ on $\partial\Omega$, then

$$G\left(\psi_*,\, \mathbf{A}_*\right) = E\left(\psi_*,\, \mathbf{A}_*\right) \tag{27}$$

and $(\psi_*, \mathbf{A}_*)$ remains a critical point of G. By the ellipticity of grad $G(\psi, \mathbf{A})$, see [7, 11, 28], we may apply the result of [27] to conclude that there exist constants $\theta_0, \sigma_1 \in (0,1)$ such that

$$|G(\psi,\, \mathbf{A}) - G(\psi_*, \mathbf{A}_*)|^{\theta_0} \le \|\text{ grad } \quad G\left(\psi,\, \mathbf{A}\right)\|_{L^2(\Omega)} \tag{28}$$

for any $\|(\psi, \mathbf{A}) - (\psi_*, \mathbf{A}_*)\|_{C^2(\Omega)} \le \sigma_1$.

Let χ be a gauge transformation function and $\tilde{\psi} = e^{i\kappa\chi}\psi$ with $\tilde{\mathbf{A}} = \mathbf{A} + \nabla\chi$. Let us choose χ such that $\operatorname{div} \tilde{\mathbf{A}} = 0$. Simple calculation shows that

$$\frac{\partial G}{\partial \tilde{\psi}} = e^{i\kappa\chi}\frac{\partial E}{\partial \psi} \qquad \text{and} \qquad \frac{\partial G}{\partial \tilde{\mathbf{A}}} = \frac{\partial E}{\partial \mathbf{A}}.$$

So,

$$\|\text{ grad } \ E(\psi, \mathbf{A})\|_{L^2(\Omega)} = \|\text{ grad } \ G(\tilde{\psi}, \tilde{\mathbf{A}})\|_{L^2(\Omega)}\ .$$

Since $\operatorname{div} \mathbf{A}_* = 0$ for small enough σ_0, if $\|(\psi, \mathbf{A}) - (\psi_*, \mathbf{A}_*)\|_{C^2(\Omega)} \le \sigma_0$, we have $\|(\tilde{\psi}, \tilde{\mathbf{A}}) - (\psi_*, \mathbf{A}_*)\|_{C^2(\Omega)} \le \sigma_1$. Thus,

$$\begin{aligned}
|E(\psi, \mathbf{A}) - E(\psi_*, \mathbf{A}_*)|^{\theta_0} &= |E(\tilde{\psi}, \tilde{\mathbf{A}}) - G(\psi_*, \mathbf{A}_*)|^{\theta_0} \\
&= |G(\tilde{\psi}, \tilde{\mathbf{A}}) - G(\psi_*, \mathbf{A}_*)|^{\theta_0} \\
&\le \|\text{ grad } \quad G\,(\tilde{\psi},\, \tilde{\mathbf{A}})\|_{L^2(\Omega)} \\
&= \|\text{ grad } \quad E(\psi, \mathbf{A})\|_{L^2(\Omega).}
\end{aligned}$$

This proves the lemma.

We now turn to the proof of Theorem 2.1. It is important to observe that both the energy functional and inequality in Lemma 2.3 are gauge invariant.

By Lemma 2.2, given any $\epsilon > 0$, there exists a sufficiently large time t_n such that

$$\|\dot{\psi}(t_n)\|_{L^2(\Omega)} < \epsilon, \tag{29}$$

$$\|\dot{\mathbf{A}}(t_n)\|_{L^2(\Omega)} < \epsilon, \tag{30}$$

and

$$\int_{t_n}^{T} \left[\|\dot{\psi}(s)\|^2_{L^2(\Omega)} + \|\dot{\mathbf{A}}(s)\|^2_{L^2(\Omega)}\right] ds < \epsilon \quad \forall\, T > t_n. \tag{31}$$

By gauge invariance, one may define a gauge transformation function $\tilde{\chi}(t_n)$ such that $\tilde{\psi}(t_n) = e^{i\kappa\tilde{\chi}(t_n)}\psi(t_n)$ and $\tilde{\mathbf{A}}(t_n) = \mathbf{A}(t_n) + \nabla\tilde{\chi}(t_n)$ and

$$\operatorname{div} \tilde{\mathbf{A}}(t_n) = 0 \quad \text{in } \Omega,$$

$$\tilde{\mathbf{A}}(t_n) \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega\ .$$

We note that $\tilde{\psi}(t) = e^{i\kappa\tilde{\chi}(t_n)}\psi(t)$ and $\tilde{\mathbf{A}}(t) = \mathbf{A}(t) + \nabla\tilde{\chi}(t_n)$ are again solutions of the time-dependent G–L equations in the zero electric potential gauge with properly modified initial conditions and equation (24) still holds. Also,

$$\|\dot{\tilde{\psi}}(t_n)\|_{L^2(\Omega)} < \epsilon \quad \text{and} \quad \|\dot{\tilde{\mathbf{A}}}(t_n)\|_{L^2(\Omega)} < \epsilon$$

and

$$\int_{t_n}^{T} \left[\|\dot{\tilde{\psi}}(s)\|^2_{L^2(\Omega)} + \|\dot{\tilde{\mathbf{A}}}(s)\|^2_{L^2(\Omega)}\right] ds < \epsilon \quad \forall\, T > t_n$$

remain valid. To apply Lemma 2.3, we must construct solutions that are $C^2(\Omega)$ close to some steady state for some time $\tilde{t_n} \in [t_n, t_n + 1]$. To get the C^2 closeness, we use another gauge transformation $\bar{\psi}(t) = e^{i\kappa\bar{\chi}(t)}\tilde{\psi}(t)$ and $\bar{\mathbf{A}}(t) = \tilde{\mathbf{A}}(t) + \nabla\bar{\chi}(t)$, where $\bar{\chi}$ is defined by

$$\frac{\partial\bar{\chi}}{\partial t} - \Delta\bar{\chi} = \operatorname{div}\tilde{\mathbf{A}} \quad \text{in} \;\; \Omega$$

with boundary condition

$$\frac{\partial\bar{\chi}}{\partial\mathbf{n}} = -\tilde{\mathbf{A}}\cdot\mathbf{n} = 0 \quad \text{on } \partial\Omega$$

and initial condition at $t = t_n$: $\bar{\chi}(t_n) = 0$.

Note that from the G–L equations (7)–(9), we may get [11]

$$\operatorname{div}\dot{\tilde{\mathbf{A}}} = -\frac{i}{2}\kappa\left[\tilde{\psi}\frac{\partial\psi^*}{\partial t} - \tilde{\psi}^*\frac{\partial\psi}{\partial t}\right].$$

Thus, we have

$$\int_{t_n}^{t_n+1}\left[\|\dot{\tilde{\mathbf{A}}}(s)\|^2_{L^2(\Omega)}\right] ds < c\epsilon$$

for some generic constant c. From this, we may get

$$\int_{t_n}^{t_n+1}\left[\|\nabla\dot{\bar{\chi}}(s)\|^2_{L^2(\Omega)}\right] ds < c\epsilon.$$

This further implies that

$$\int_{t_n}^{t_n+1}\left[\|\Delta\bar{\chi}(s)\|^2_{L^2(\Omega)} + \|\dot{\bar{\chi}}(s)\|^2_{L^2(\Omega)}\right] ds < c\epsilon.$$

It follows that $(\bar{\psi}, \bar{\mathbf{A}})$ also satisfies estimates similar to those in equations (29)–(31). Notice that $(\bar{\psi}, \bar{\mathbf{A}})$ is a solution of the time-dependent G–L equations in the gauge $\Phi = -\operatorname{div}\mathbf{A}$ and it satisfies a parabolic system:

$$(32) \qquad \left\{ \begin{aligned} &\frac{\partial\bar{\psi}}{\partial t} - i\,\kappa\operatorname{div}\bar{\mathbf{A}}\,\bar{\psi} + \left(\frac{i}{\kappa}\nabla + \bar{\mathbf{A}}\right)^2\bar{\psi} - \bar{\psi} + |\bar{\psi}|^2\,\bar{\psi} = 0, \\ &\frac{\partial\bar{\mathbf{A}}}{\partial t} - \Delta\bar{\mathbf{A}} = -\frac{i}{2\kappa}\left(\bar{\psi}^*\,\nabla\bar{\psi} - \bar{\psi}\,\nabla\bar{\psi}^*\right) - \bar{\mathbf{A}}|\bar{\psi}|^2 \end{aligned} \right.$$

with boundary conditions

$$\operatorname{curl} \bar{\mathbf{A}} \wedge \mathbf{n} = \mathbf{H}_0 \wedge \mathbf{n} \,, \quad \bar{\mathbf{A}} \cdot \mathbf{n} = 0, \quad \text{and} \quad \nabla \bar{\psi} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega.$$

Using standard parabolic regularity and estimates like (29)–(31) for $(\bar{\psi}, \bar{\mathbf{A}})$, we may find a small $\delta_0 > 0$ such that for all $\delta_0 < \delta < 1$, we have

$$\|\dot{\bar{\psi}}(t_n+\delta)\|_{C^2(\Omega)} < c\epsilon \quad \text{and} \quad \|\dot{\bar{\mathbf{A}}}(t_n+\delta)\|_{C^2(\Omega)} < c\epsilon$$

and

$$\|\operatorname{div} \bar{\mathbf{A}}(t_n+\delta)\|_{C^1(\Omega)} < c\epsilon$$

for some constant c. Thus, we may conclude that in $C^2(\Omega)$, $(\bar{\psi}(t_n+\delta), \bar{\mathbf{A}}(t_n+\delta))$ is close to some solution $(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)$ of the steady state G–L equations with $\operatorname{div} \bar{\mathbf{A}}_\infty = 0$. Now, by Lemma 2.3, we get

$$\begin{aligned} &\left|E(\bar{\psi}(t_n+\delta), \bar{\mathbf{A}}(t_n+\delta)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)\right|^{\theta_0} \\ &\quad \le \| \text{ grad } \quad E(\bar{\psi}(t_n+\delta), \bar{\mathbf{A}}(t_n+\delta))\|_{L^2(\Omega)} \,. \end{aligned}$$

Using the gauge invariance, however, this also implies that

$$\begin{aligned} &\left|E(\psi(t_n+\delta), \mathbf{A}(t_n+\delta)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)\right|^{\theta_0} \\ &\quad \le \| \text{ grad } \quad E(\psi(t_n+\delta), \mathbf{A}(t_n+\delta))\|_{L^2(\Omega)} \,. \end{aligned}$$

Since $(\bar{\psi}(t_n+\delta), \bar{\mathbf{A}}(t_n+\delta))$ is C^2 close to $(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)$, we have

$$\begin{aligned} &\left|E(\psi(t_n+\delta), \mathbf{A}(t_n+\delta)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)\right| \\ &\quad = \left|E(\bar{\psi}(t_n+\delta), \bar{\mathbf{A}}(t_n+\delta)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty)\right| \\ &\quad \le c\epsilon \,. \end{aligned}$$

By the monotonicity of $E(\psi(t), \mathbf{A}(t))$, we have that for any T, there exists t_n large enough such that

$$\begin{aligned} &E(\psi(T), \mathbf{A}(T)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty) \\ &\quad \ge E(\psi(t_n+\delta), \mathbf{A}(t_n+\delta)) - E(\bar{\psi}_\infty, \bar{\mathbf{A}}_\infty) \\ &\quad \ge c\epsilon \end{aligned}$$

for some generic constant c. Using the ideas given in Case (iii) for ODEs, we conclude that

$$\int_T^\infty \left[\|\dot{\bar{\psi}}(s)\|_{L^2(\Omega)} + \|\dot{\bar{\mathbf{A}}}(s)\|_{L^2(\Omega)}\right] ds < c\epsilon.$$

Thus, we must have that

$$V_\infty = \lim_{t\to\infty} \; v(t) \quad \text{exists in } \mathbf{V}.$$

This concludes the proof of the Theorem 2.1.

3. The pinning effect of variable thickness in thin films. In the previous section, we were concerned with the dynamic properties of solutions to the time-dependent G–L equations. In this section, we focus our attention on the static case. In particular, we illustrate that the vortices may be pinned by inhomogeneities inside the material. The inhomogeneities we consider here are introduced due to the variation in thickness of the sample. During our writing, we also became aware of independent works [9, 1, 18] on the same subject; thus, we only give a brief outline here of the approach we have used.

3.1. The Dirichlet case. To present the main idea, we first ignore the magnetic potential and consider

$$\begin{aligned} &\min\{\mathcal{F}^a_\epsilon(\psi)\ ,\ \psi|_{\partial\Omega}=g\ \} \\ &\quad = \min\left\{ \int_\Omega a(\mathbf{x}) \left(\frac{1}{2}\left|\nabla\psi\right|^2 + \frac{1}{4\epsilon^2}(1-|\psi|^2)^2 \right) d\mathbf{x}\ ,\ \ \psi|_{\partial\Omega}=g\ \right\}\ . \end{aligned} \tag{33}$$

For convenience, we assume that Ω is a bounded, smooth (say Lipschitz) domain in $\mathbb{R}^2$. We may also view ψ as a map from Ω to $\mathbb{R}^2$ and g as a smooth map from $\partial\Omega$ to $\mathbf{S}^1$ with $deg(g,\partial\Omega)=d(\geq 0)$. The coefficient a is a smooth (say Lipschitz or Hölder) function from Ω to $\mathbb{R}^R$ with $0<m\leq a(\mathbf{x})\leq M$ for all $\mathbf{x}\in\bar{\Omega}$. The minimizers of functional $\mathcal{F}^a_\epsilon$ satisfy

$$\operatorname{div}\left(a(\mathbf{x})\nabla\psi(\mathbf{x})\right) + \frac{a(\mathbf{x})}{\epsilon^2}(1-|\psi(\mathbf{x})|^2)\psi(\mathbf{x}) = 0 \quad \text{in } \Omega. \tag{34}$$

First of all, let us consider the case where, in Ω, there are at least d distinct points $\mathbf{b}_1,\ldots,\mathbf{b}_d,\ldots,\mathbf{b}_k$ with $a(\mathbf{b}_j)=m, j=1,2,\ldots,k$. Moreover, we assume that each $\mathbf{b}_j$ is a strict minimum, that is, $a(\mathbf{x})>m$ for any $\mathbf{x}\neq\mathbf{b}_j$ in a small neighborhood of $\mathbf{b}_j$. We define

$$\delta_0 = \frac{1}{2}\min\{|\mathbf{b}_j-\mathbf{b}_i|, dist(\mathbf{b}_j,\partial\Omega); j\neq i\ ,\ i,j=1,\ldots,k\} > 0. \tag{35}$$

THEOREM 3.1. *Let ψ_{ϵ_n}, $\epsilon_n \searrow 0$, be a sequence of minimizers of* (33). *Then*

$$\psi_{\epsilon_n}(\mathbf{x}) \to \psi^*(\mathbf{x}) \qquad in\ C^{1,\alpha}_{\text{loc}}\left(\overline{\Omega}/\{\mathbf{b}_1,\ldots,\mathbf{b}_d\}\right), \tag{36}$$

where

$$\psi^*(\mathbf{x}) = \prod_{j=1}^{d} \frac{\mathbf{x}-\mathbf{b}_j}{|\mathbf{x}-\mathbf{b}_j|} e^{ih^*(\mathbf{x})} \qquad in\ \Omega \tag{37}$$

and $\mathbf{b}_1,\ldots,\mathbf{b}_d$ are d distinct points in Ω with

$$a(\mathbf{b}_j) = m = \min_{\mathbf{x}\in\bar{\Omega}} a(\mathbf{x})\ . \tag{38}$$

Moreover, if we write $\psi^*(\mathbf{x}) = e^{i(\Theta(\mathbf{x})+h^*(\mathbf{x}))}$, then $h^*\in H^1(\Omega)\cap C^\alpha(\bar{\Omega})$, $\psi^*=g$ on $\partial\Omega$ and

$$\operatorname{div}\ [a(\mathbf{x})(\nabla\Theta+\nabla h^*)] = 0 \qquad in\ \ \Omega/\{\mathbf{b}_1,\ldots,\mathbf{b}_d\}. \tag{39}$$

The proof of the above theorem is based on a series of estimates as similarly presented in an earlier work [20]. One first has the energy upper bound

$$\min\{\mathcal{F}_\epsilon^a(\psi),\ \ \psi|_{\partial\Omega}=g\ \}\le m\pi d\log\frac{1}{\epsilon}+C(a,g,\Omega)$$

and energy lower bound

$$\min\left\{\int_B a(\mathbf{x})\left(\frac{1}{2}|\nabla\psi|^2+\frac{1}{4\epsilon^2}(1-|\psi|^2)^2\right)d\mathbf{x},\ \ \psi|_{\partial B}=g\ ,\ |g|\le 1\right\}$$
$$\ge m\pi d\log\frac{1}{\epsilon}-C(K), \tag{40}$$

where $m=\min_{\mathbf{x}\in B}a(\mathbf{x})$ if $deg(g,\partial B)\neq 0$. Now, the class $S_g(\lambda,K)$ may be defined as in [20].

Definition 3.2. *Let Ω, g be given as before. We say a map $u:\Omega\to\mathbb{R}^2$ belongs to the class $S_g(\lambda,K)$ if*

(i) $\mathcal{F}_\epsilon(u)\le m\pi d\log\frac{1}{\epsilon}+K$.

(ii) *for any* $\mathbf{x}\in\{\mathbf{x}\in\Omega,|u(\mathbf{x})|\le\frac{1}{2}\}$, $B_{\lambda\epsilon}(\mathbf{x})\cap\Omega\subset\{|u(\mathbf{x})|\le\frac{3}{4}\}$.

Then one can show that if ψ minimizes (33), there exists some positive constants λ,K such that $\psi\in S_g(\lambda,K)$.

One may then show that there are exactly d balls, say $B_1,\dots,B_d$, with $\mathbf{x}_1^\epsilon,\dots,\mathbf{x}_d^\epsilon$ their centers, such that the corresponding $d_j=deg(u,\partial B_j)=1$ and

$$\min\{|\mathbf{x}_j^\epsilon-\mathbf{x}_k^\epsilon|,dist(\mathbf{x}_j^\epsilon,\partial\Omega);j,k=1,\dots,d,j\neq k\ \}\ge\delta_1(\lambda,K)>0.$$

After extracting a subsequence, we have

$$\mathbf{x}_j^{\epsilon_n}\to\mathbf{x}_j^*\quad\text{as }\epsilon_n\to 0.$$

The limits $\{\mathbf{x}_j^*\}$ are all different points; moreover, $a(\mathbf{x}_j^*)=m,\ j=1,2,\dots,d$. Combining with estimates away from vortices, the proof now follows similarly to that in [3, 20].

In the following, we briefly discuss the case when the number of minima is less than the degree of the boundary data. For simplicity, we focus on the case where Ω is the unit disc B in $\mathbb{R}^2$ and $g(\theta)=e^{id\theta}$ for some positive integer $d\ge 2$. Let $a(x,y)=1+r^2,r^2=x^2+y^2$. We consider

$$E_B=\min\left\{\mathcal{F}_{\epsilon,B}^a(\psi)=\int_B\left(\frac{a}{2}|\nabla\psi|^2+\frac{a}{4\epsilon^2}(1-|\psi|^2)^2\right)dxdy,\right.$$
$$\left.\psi|_{\partial\Omega}=g\right\}. \tag{41}$$

Theorem 3.3. *Let ψ_ϵ be a sequence of minimizers of* (41) *as $\epsilon\to 0$. Then all vortices of the ψ must be separated and of degree $+1$ for small enough $\epsilon>0$ and they go to the origin as $\epsilon\to 0$.*

Proof. Indeed, for any small parameter $\delta>0$ (say $\delta<1/d$), one may choose $\epsilon\le\epsilon(\delta)$ and ψ with $\psi|_{\partial\Omega}=g$ and vortices placed along $\partial B_{\delta/2}$ such that

$$E_B\le\mathcal{F}_{\epsilon,B}^a(\psi)\le\pi d(1+\delta^2)\log\frac{\delta}{\epsilon}+\pi d^2\log\frac{1}{\delta}+c_0d^2$$
$$=\pi d(1+\delta^2)\log\frac{1}{\epsilon}+\{\pi d^2-\pi d(1+\delta^2)\}\log\frac{1}{\delta}+c_0d^2\ .$$

On the other hand, if there are either vortices of degree no less than 2 or vortices outside $B_{\sqrt{\delta}}(0)$, then by [3],

$$\mathcal{F}^a_{\epsilon,B}(\psi) \geq \pi(d + d\delta)\log\frac{1}{\epsilon} + c(\delta, d)\ .$$

For $\delta > 0$ and $\epsilon \to 0$, since $\pi(d + d\delta)\log\frac{1}{\epsilon} + c(\delta, d)$, one sees that all vortices of the minimizer must go to the origin and be of degree $+1$.

The next question is for given d (say $2 \leq d \leq 5$) and $\epsilon > 0$ (but small): How close must these vortices be to the origin? We want to show that they cannot be too close.

Indeed, for a small generic constant β, if all vortices of ψ are in the β-neighborhood of the origin, then

$$\int_{B_\beta(0)} (1 + r^2)\left(\frac{1}{2}|\nabla\psi|^2 + \frac{1}{4\epsilon^2}(1 - |\psi|^2)^2\right) dxdy \geq \pi d \log\frac{\beta}{\epsilon} - c_0 d^2$$

and

$$\int_{B\setminus B_\beta(0)} (1 + r^2)\left(\frac{1}{2}|\nabla\psi|^2 + \frac{1}{4\epsilon^2}(1 - |\psi|^2)^2\right) dxdy \geq \pi d^2 \log\frac{1}{\beta}\ .$$

So,

$$\mathcal{F}^a_{\epsilon,B}(\psi) \geq \pi d \log\frac{1}{\epsilon} + \pi d(d-1)\log\frac{1}{\beta} - c_0 d^2.$$

On the other hand, by placing vortices along $\partial B_{\delta/2}(0)$, one may construct a map with energy

$$\mathcal{F}^a_{\epsilon,B}(\psi) \leq \pi d(1 + \delta^2)\log\frac{1}{\epsilon} + \pi d(d - 1 - \delta^2)\log\frac{1}{\delta} + c_0 d^2\ .$$

Comparing the two bounds, we see that not all vortices are in the δ-neighborhood of the origin if we have

$$\pi d(d-1)\log\frac{\delta}{\beta} \geq 2c_0 d^2 + \pi d\delta^2 \log\frac{1}{\epsilon}\ .$$

By taking $\beta \leq \delta^2$, this means

$$\pi d(d-1)\log\frac{1}{\delta} - \pi d\delta^2\log\frac{1}{\epsilon} \geq 2c_0 d^2\ .$$

Again, letting $\log\frac{1}{\delta} \geq 2c_0$, it is sufficient to have

$$\frac{d-1}{2}\log\frac{1}{\delta} \geq \delta^2 \log\frac{1}{\epsilon}\ ,$$

or

$$\epsilon \geq \delta^{\frac{d-1}{2\delta^2}}\ .$$

Thus, for small but positive ϵ, δ cannot be too small.

We have performed a series of numerical experiments to illustrate the pinning effect even for modest values of ϵ. The numerical methods used here are similar to

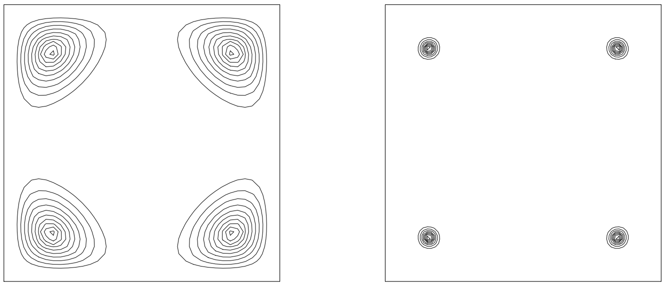


Fig. 2. *Contour plots of the magnitude of the order parameter for a model with constant thickness. The left-hand figure is for $\epsilon = .1$. The right-hand figure is for $\epsilon = .02$.*

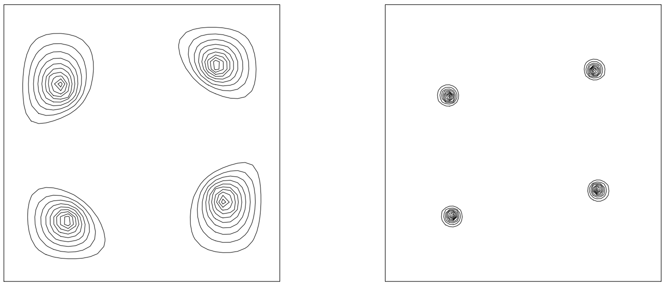


Fig. 3. *Contour plots of the magnitude of the order parameter for a model with variable thickness $a = a(x, y)$. The left-hand figure is for $\epsilon = .1$. The right-hand figure is for $\epsilon = .02$.*

those discussed in [13, 14]. For computational convenience, the unit square $[0, 1]^2$ is taken to be our sample Ω. In the first experiment, we choose the thickness function $a \equiv 1$. We solve for the minimizer of (33) by using the Dirichlet boundary conditions with $|\psi| = 1$ on the boundary. $\psi|_{\partial\Omega}$ has a winding number 4. The contour plots of the magnitude of the order parameter ψ are given in Figure 2.

Next, we choose the thickness function $a = a(x, y)$ such that it has four minima at $(0.25, 0.25), (0.75, 0.35), (0.25, 0.65), (0.75, 0.75)$ with the same minimum value. The contour plots are given in Figure 3. We see that as ϵ gets smaller, the vortices get "pinned" at the minima of a. This is the case illustrated by Theorem 3.1.

Now, we continue the numerical experiments again using the Dirichlet boundary conditions with $|\psi| = 1$ on the boundary. However, we impose the boundary condition such that $\psi|_{\partial\Omega}$ has a winding number 3. We first choose the thickness function $a = a(x, y)$ such that it has four minima at $(0.25, 0.25), (0.25, 0.65), (0.75, 0.35), (0.75, 0.75)$ with the same minimum value. The contour plots are given in Figure 4. Since the number of minima is larger than the winding number, each vortex gets pinned to a different minimum point.

Then, we change the thickness function $a = a(x, y)$ such that it has two minima at $(0.25, 0.65), (0.75, 0.35)$ with the same minimum value. The contour plots are given in Figure 5. Since the number of minima is less than the winding number, one vortex gets pinned at one minimum but the other two get attracted to another minimum with some distance still between them.

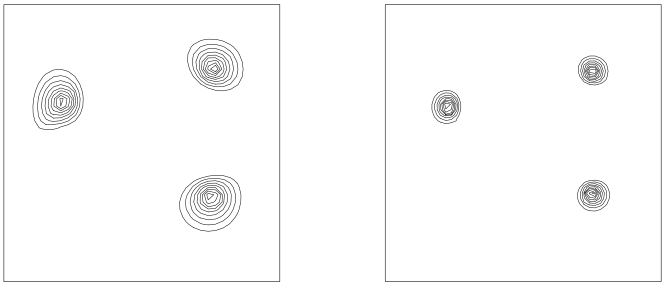


FIG. 4. *Contour plots of the magnitude of the order parameter for a model with variable thickness $a = a(x, y)$. The left-hand figure is for $\epsilon = .08$. The right-hand figure is for $\epsilon = .04$.*

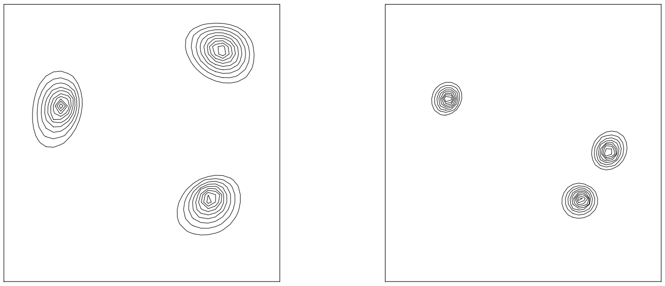


FIG. 5. *Contour plots of the magnitude of the order parameter for a model with variable thickness $a = a(x, y)$. The left-hand figure is for $\epsilon = .08$. The right-hand figure is for $\epsilon = .04$.*

For comparison purposes, we also give the pictures when choosing the thickness function $a \equiv 1$. The contour plots are given in Figure 6.

Finally, we present an experiment on the co-existence of vortex–antivortex solutions to the minimizers of (33) with constant thickness functions. This question has been rigorously studied in [22]. We again use the Dirichlet boundary conditions with $|\psi| = 1$ on the boundary. $\psi|_{\partial\Omega}$ has a winding number 0. Here, we start with a vortex state that has a vortex with winding number $+1$ on one side of the domain and a vortex with winding number -1 on the other side of the domain. We numerically follow the gradient flow of (33). The solution reaches steady state which again consists of two vortices with opposite winding numbers. The contour plots of the magnitude of the steady state solution ψ are given in Figure 7.

3.2. The Neumann problem. We now study the minimization problem with Neumann-type boundary conditions. Let $a(\mathbf{x})$ have a strict local minimum at points $\mathbf{b}_1, \ldots, \mathbf{b}_d$ with $a(\mathbf{b}_j) = m$, $j = 1, \ldots, d$. Recall that

$$\delta_0 = \frac{1}{2} \min\{|\mathbf{b}_j - \mathbf{b}_k|, dist(\mathbf{b}_j, \partial\Omega); j \neq k, \;\; j, k = 1, \ldots, d\} > 0.$$

Let $r_0 < \delta_0$, $\Omega_0 = \Omega \setminus \bigcup_{i=1}^{d} B_{r_0}(\mathbf{b}_i)$ and define

$$V = \Bigg\{ u \in H^1(\Omega) \mid \;\; |u| \geq \frac{1}{2} \;\text{ in } \Omega_0,$$

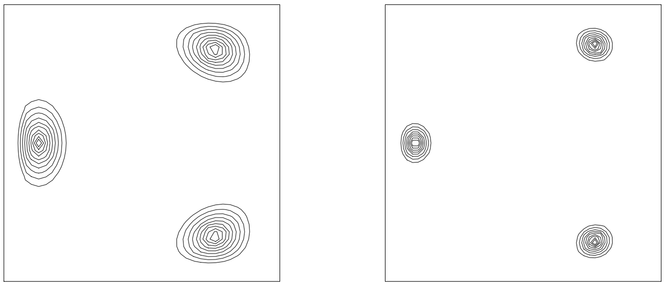


FIG. 6. *Contour plots of the magnitude of the order parameter for a model with constant thickness. The left-hand figure is for* $\epsilon = .08$. *The right-hand figure is for* $\epsilon = .04$.

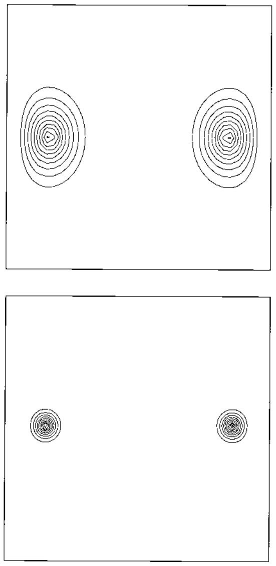


FIG. 7. *Contour plots of the magnitude of the order parameter. The top figure is for* $\epsilon = .08$. *The bottom figure is for* $\epsilon = .03$.

$$deg\left(\frac{u}{|u|}, \partial B_{r_0}(\mathbf{b}_k)\right) = 1\ ,\ 1 \le k \le d.\Big\}\ . \tag{42}$$

We consider

$$\begin{aligned} &\min\{\mathcal{F}^a_\epsilon(\psi)\ ,\ \psi \in V\ \} \\ &= \min\left\{\int_\Omega a(\mathbf{x})\left(\frac{1}{2}\left|\nabla\psi\right|^2 + \frac{1}{4\epsilon^2}(1-|\psi|^2)^2\right)d\mathbf{x}\ ,\ \ \psi \in V\ \right\}\ . \end{aligned} \tag{43}$$

It is easy to see that V is a weakly closed subset of $H^1(\Omega)$. In fact, it is also connected. For given $\epsilon > 0$, (43) has at least one minimizer, denoted by ψ_ϵ.

THEOREM 3.4. *Let* ψ_{ϵ_n}, $\epsilon_n \searrow 0$, *be a sequence of minimizers of* (43). *Then,*

$$\psi_{\epsilon_n}(\mathbf{x}) \to \psi^*(\mathbf{x}) \qquad in\ C^{1,\alpha}_{\text{loc}}\left(\overline{\Omega}/\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}\right),$$

where

$$\psi^*(\mathbf{x}) = \prod_{j=1}^{d} \frac{\mathbf{x}-\mathbf{b}_j}{|\mathbf{x}-\mathbf{b}_j|} e^{ih^*(\mathbf{x})} \qquad \text{in } \Omega.$$

Moreover, if we write $\psi^*(\mathbf{x}) = e^{i(\Theta(\mathbf{x})+h^*(\mathbf{x}))}$, *then* $h^* \in H^1(\Omega) \cap C^\alpha(\bar{\Omega})$, $\frac{\partial \phi^*}{\partial n} = 0$ *on* $\partial\Omega$ *and*

$$\operatorname{div}\,[a(\mathbf{x})(\nabla\Theta + \nabla h^*)] = 0 \qquad \text{in} \quad \Omega/\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}.$$

The proof follows from constructions similar to those outlined in section 3.1.

3.3. Problems with prescribed magnetic potential. Unlike bulk material, in the thin film limit, superconductors of type-I and type-II display vortex-like structure. It has been shown that the G–L functional takes the special form

$$\mathcal{F}(\psi) = \int_\Omega a(\mathbf{x}) \left(\frac{1}{2} \left| i\nabla\psi + A_0(\mathbf{x})\psi \right|^2 + (1-|\psi|^2)^2 \right) d\mathbf{x},$$

where $a(\mathbf{x})$ represents the relative thickness distribution of the thin film. Since, to leading order, the magnetic field penetrates the film uniformly, we get A_0 to be a given magnetic potential that can be prescribed by setting $\operatorname{curl} A_0(\mathbf{x}) = H$.

Numerical simulation in [6, 14] suggests that the thickness variation provides a pinning mechanism for vortices. Using techniques similar to those in section 3.1, we now provide a rigorous analysis for such phenomena in the case where κ is large while the magnetic field is weak ($H \approx \kappa^{-1}$). By a proper scaling of the free energy, we get a functional of the form

$$\mathcal{F}^a_\epsilon(\psi) = \int_\Omega a(\mathbf{x}) \left(\frac{1}{2} \left| \nabla\psi - iA_0(\mathbf{x})\psi \right|^2 + \frac{1}{4\epsilon^2}(1-|\psi|^2)^2 \right) d\mathbf{x}. \tag{44}$$

In the present form, ϵ is a small parameter measuring the relative penetration depth and the sample size. Hence, we may consider problems similar to those in section 3.1 with the functional defined above. We first consider the Dirichlet problem

$$\begin{aligned} \min\left\{\mathcal{F}^a_\epsilon(\psi)\ ,\ \psi|_{\partial\Omega} = g\ \right\} = \min\Bigg\{ \int_\Omega a(\mathbf{x}) \Bigg(& \frac{1}{2} \left| \nabla\psi - A_0(\mathbf{x})\psi \right|^2 \\ & + \frac{1}{4\epsilon^2}(1-|\psi|^2)^2 \Bigg) d\mathbf{x}\ ,\ \ \psi|_{\partial\Omega} = g\ \Bigg\}\ . \end{aligned} \tag{45}$$

For simplicity, we assume that $|g| = 1$ on $\partial\Omega$, $deg(g, \partial\Omega) = d$, and $a(\mathbf{x})$ has d distinct strict minimums at points $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d$. With a proper choice of gauge, we may define

$$A_0(x,y) = \frac{1}{2}(Hy, -Hx)^T,$$

where H is the scaled applied magnetic field. We then have the following theorem.

Theorem 3.5. *Let* ψ_{ϵ_n}, $\epsilon_n \searrow 0$, *be a sequence of minimizers of* (45). *Then*

$$\psi_{\epsilon_n}(\mathbf{x}) \to \psi^*(\mathbf{x}) \qquad \text{in } C^{1,\alpha}_{\text{loc}}\left(\overline{\Omega}/\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}\right),$$

where

$$\psi^*(\mathbf{x}) = \prod_{j=1}^{d} \frac{\mathbf{x}-\mathbf{b}_j}{|\mathbf{x}-\mathbf{b}_j|} e^{ih^*(\mathbf{x})} \qquad \text{in } \Omega.$$

Moreover, if we write $\psi^*(\mathbf{x}) = e^{i(\Theta(\mathbf{x})+h^*(\mathbf{x}))}$, *then* $h^* \in H^1(\Omega)\cap C^\alpha(\bar{\Omega})$, $\psi^*|_{\partial\Omega} = g$ *on* $\partial\Omega$, *and*

$$\operatorname{div}\,[a(\mathbf{x})(\nabla\Theta + \nabla h^*)] = 0 \qquad \text{in} \quad \Omega/\{\mathbf{b}_1,\ldots,\mathbf{b}_d\}.$$

Next, we may also prove similar results for the Neumann-type problems. Let V be the space defined in (42). We consider

$$\begin{aligned}\min\{\mathcal{F}_\epsilon^a(\psi)\ ,\ \psi\in V\ \} = \min\Bigg\{ \int_\Omega a(\mathbf{x}) \bigg(&\frac{1}{2}\left|\nabla\psi - A_0(\mathbf{x})\psi\right|^2 \\ &+\frac{1}{4\epsilon^2}(1-|\psi|^2)^2\bigg)\,d\mathbf{x}\ ,\quad \psi\in V\ \Bigg\}.\end{aligned} \tag{46}$$

Theorem 3.6. *Let* ψ_{ϵ_n}, $\epsilon_n \searrow 0$, *be a sequence of minimizers of* (46). *Then*

$$\psi_{\epsilon_n}(\mathbf{x}) \to \psi^*(\mathbf{x}) \qquad \text{in } C^{1,\alpha}_{\text{loc}}\left(\overline{\Omega}/\{\mathbf{b}_1,\ldots,\mathbf{b}_d\}\right),$$

where

$$\psi^*(\mathbf{x}) = \prod_{j=1}^{d} \frac{\mathbf{x}-\mathbf{b}_j}{|\mathbf{x}-\mathbf{b}_j|} e^{ih^*(\mathbf{x})} \qquad \text{in } \Omega.$$

Moreover, if we write $\psi^*(\mathbf{x}) = e^{i(\Theta(\mathbf{x})+h^*(\mathbf{x}))} = e^{i(\phi*(\mathbf{x}))}$, *then* $h^*\in H^1(\Omega)\cap C^\alpha(\bar{\Omega})$, $\frac{\partial\phi^*}{\partial n} = A_0\cdot\mathbf{n}$ *on* $\partial\Omega$ *(here,* $\mathbf{n}$ *is the outward normal of* $\partial\Omega$*), and*

$$\operatorname{div}\,[a(\mathbf{x})(\nabla\Theta + \nabla h^*)] = 0 \qquad \text{in} \quad \Omega/\{\mathbf{b}_1,\ldots,\mathbf{b}_d\}.$$

Again, the proofs of Theorems 3.3 and 3.4 are omitted due to their similarities to our earlier discussions.

4. The renormalized energy and the vortex solution of the full G–L model with applied magnetic field. With proper scaling, we focus on the following form of the G–L functional:

$$\mathcal{G}(\psi,\mathbf{A}) = \int_\Omega \left(\frac{1}{4\epsilon^2}(1-|\psi|^2)^2 + \frac{1}{2}\left|(\nabla - i\mathbf{A})\,\psi\right|^2 + \frac{1}{2}|\operatorname{curl}\mathbf{A} - H_0|^2 \right) d\mathbf{x}.$$

Let $\Omega \in \mathbb{R}^2$ be a bounded Lipshitz domain and H_0 be a constant field. In this nondimensionalization, one may view ϵ as proportional to $\frac{1}{\kappa}$ and H_0 as proportional to κ times the (nondimensionalized) applied field. Studies of the densely packed vortex state when the nondimensionalized field is near the upper critical field may be found in [4]. In the low field case, variational problems concerning the G–L functional have been considered in [5] but with boundary conditions that are not completely physical. Our discussion here is valid for the natural (physically meaningful) boundary conditions and we will borrow many useful results from [5] and [20].

4.1. The renormalized energy. Following the discussions in [20, 21, 5], we now formulate the renormalized energy: Let

$$\psi = e^{i\phi_b(\mathbf{x})} = \prod_{j=1}^{d} \frac{\mathbf{x} - \mathbf{b}_j}{|\mathbf{x} - \mathbf{b}_j|} e^{ih}$$

for some points $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d) \in \Omega^d$ and $\frac{\partial \phi_b}{\partial \mathbf{n}} = 0$ on $\partial\Omega$. Let

$$B_\rho = \bigcup_{j=1}^{d} B_\rho(\mathbf{b}_j) \ .$$

Choose the gauge $\operatorname{div} \mathbf{A} = 0$ in Ω and $\mathbf{A} \cdot \mathbf{n} = 0$ on $\partial\Omega$. We may define ζ such that

$$\mathbf{A} = \nabla^\perp \zeta \quad \text{in } \ \Omega,$$

$$\zeta = 0 \quad \text{on } \ \partial\Omega.$$

Now, consider

$$\begin{aligned}
\mathcal{G}_\rho \ &= \int_{\Omega \setminus B_\rho} \left(\frac{1}{2} \left| \nabla \phi_b - \nabla^\perp \zeta \right|^2 + |\Delta \zeta - H_0|^2 \right) d\Omega \\
&= \int_{\Omega \setminus B_\rho} \left(\frac{1}{2} |\nabla \phi_b|^2 + \frac{1}{2} |\nabla \zeta|^2 - \nabla \phi_b \cdot \nabla^\perp \zeta + \frac{1}{2} \left| \Delta \zeta - H_0 \right|^2 \right) d\Omega \\
&:= d\pi \log \frac{1}{\rho} + W_\Omega(\mathbf{b}, H_0) + O(\rho),
\end{aligned}$$

where the last equality may be taken as the definition of the renormalized energy $W_\Omega(\mathbf{b}, H_0)$. Note

$$\begin{aligned}
-\int_{\Omega \setminus B_\rho)} \nabla \phi_b \cdot \nabla^\perp \zeta d\Omega &= \int_{\Omega \setminus B_\rho} \operatorname{div} (\zeta \cdot \nabla^\perp \phi_b) d\Omega \\
&= \sum_{j=1}^{d} 2\pi \zeta(\mathbf{b}_j) + O(\rho).
\end{aligned}$$

So,

$$\begin{aligned}
W_\Omega(\mathbf{b}, H_0) = & \int_{\Omega \setminus B_\rho} \frac{1}{2} |\nabla \phi_b|^2 d\Omega - d\pi \log \frac{1}{\rho} + \sum_{j=1}^{d} 2\pi \zeta(\mathbf{b}_j) \\
& + \int_{\Omega \setminus B_\rho} \frac{1}{2} \left(|\nabla \zeta|^2 + |\Delta \zeta - H_0|^2 \right) d\Omega + O(\rho) \\
= & \int_{\Omega \setminus B_\rho} \frac{1}{2} |\nabla \phi_b|^2 d\Omega - d\pi \log \frac{1}{\rho} + 2\pi \sum_{j=1}^{d} \zeta(\mathbf{b}_j) \\
& + \int_{\Omega \setminus B_\rho} \frac{1}{2} \left(|\nabla \zeta|^2 + |\Delta \zeta|^2 \right) d\Omega + \frac{1}{2} H_0^2 |\Omega| \\
& - \int_{\Omega \setminus B_\rho} H_0 \Delta \zeta d\Omega + O(\rho).
\end{aligned}$$

Minimizing the term involving ϕ_b, we see that ϕ_b is a multivalued harmonic function on $\Omega \setminus \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d\}$ and we may define the function $g_\Omega(\mathbf{b})$ by

$$\int_{\Omega\setminus B_\rho} \frac{1}{2}|\nabla\phi_b|^2 d\Omega - d\pi \log\frac{1}{\rho} := g_\Omega(\mathbf{b}) + O(\rho).$$

Minimizing the terms involving ζ, we can choose ζ to satisfy

$$-\Delta^2\zeta + \Delta\zeta = 2\pi\sum_{j=1}^{d}\delta_{\mathbf{b}_j} \quad \text{in } \Omega,$$

where $\delta_{\mathbf{b}_j}$ is the Dirac–Delta measure with boundary conditions

$$\zeta = 0 \quad \text{on } \partial\Omega,$$

$$\Delta\zeta = H_0 \quad \text{on } \partial\Omega.$$

So,

$$2\pi\sum_{j=1}^{d}\zeta(\mathbf{b}_j) = -\int_\Omega \left(|\nabla\zeta|^2 + |\Delta\zeta|^2\right) d\Omega + H_0\int_{\partial\Omega}\frac{\partial\zeta}{\partial n} d\Gamma\ .$$

Because H_0 is a constant, we get

$$\int_{\Omega\setminus B_\rho} H_0\Delta\zeta d\Omega = H_0\int_{\partial\Omega}\frac{\partial\zeta}{\partial n}d\Gamma + H_0 O(\rho)\ .$$

Thus,

$$W_\Omega(\mathbf{b}, H_0) = \frac{1}{2}H_0^2|\Omega| - \frac{1}{2}\int_\Omega \left(|\nabla\zeta|^2 + |\Delta\zeta|^2\right) d\Omega + H_0 O(\rho) + g_\Omega(\mathbf{b})\ .$$

Now, let us define $\zeta = \zeta_b + \zeta_{H_0}$, where

$$-\Delta^2\zeta_b + \Delta\zeta_b = 2\pi\sum_{j=1}^{d}\delta_{\mathbf{b}_j} \quad \text{in } \Omega,$$

$$\zeta_b = 0 \quad \text{on } \partial\Omega,$$

$$\Delta\zeta_b = 0 \quad \text{on } \partial\Omega,$$

and $\zeta_{H_0} = H_0\zeta_1$ with

$$-\Delta^2\zeta_1 + \Delta\zeta_1 = 0 \quad \text{in } \Omega,$$

$$\zeta_1 = 0 \quad \text{on } \partial\Omega,$$

$$\Delta\zeta_1 = 1 \quad \text{on } \partial\Omega.$$

Then

$$\int_\Omega |\nabla\zeta|^2 d\Omega = H_0^2 \int_\Omega |\nabla\zeta_1|^2 d\Omega + \int_\Omega |\nabla\zeta_b|^2 d\Omega - 2H_0 \int_\Omega \Delta\zeta_1 \zeta_b d\Omega,$$

$$\int_\Omega |\Delta\zeta|^2 d\Omega = H_0^2 \int_\Omega |\Delta\zeta_1|^2 d\Omega + \int_\Omega |\Delta\zeta_b|^2 d\Omega + 2H_0 \int_\Omega \Delta\zeta_1 \Delta\zeta_b d\Omega,$$

and

$$\begin{aligned}\int_\Omega \Delta\zeta_1(\Delta\zeta_b - \zeta_b) d\Omega &= \int_\Omega \zeta_1 \Delta(\Delta\zeta_b - \zeta_b) d\Omega \\ &= -2\pi \sum_{j=1}^d \zeta_1(\mathbf{b}_j)\ .\end{aligned}$$

So,

$$\begin{aligned}\int_\Omega \left(|\nabla\zeta|^2 + |\Delta\zeta|^2\right) d\Omega = H_0^2 \int_\Omega \left(|\nabla\zeta_1|^2 + |\Delta\zeta_1|^2\right) d\Omega \\ + \int_\Omega \left(|\nabla\zeta_b|^2 + |\Delta\zeta_b|^2\right) d\Omega - 2\pi \sum_{j=1}^d \zeta_1(\mathbf{b}_j)\ .\end{aligned}$$

Therefore,

$$W_\Omega(\mathbf{b}, H_0) = \frac{1}{2} H_0^2 C(\Omega) + 2\pi H_0 \sum_{j=1}^d \zeta_1(\mathbf{b}_j) + \tilde{g}_\Omega(\mathbf{b}) + O(\rho), \tag{47}$$

where $C(\Omega)$ is a constant and $\tilde{g}_\Omega(\mathbf{b})$ has the property

$$\tilde{g}_\Omega(\mathbf{b}) = \begin{cases} +\infty\ , & \mathbf{b}_i = \mathbf{b}_j \quad \text{for some } i \neq j\ , \\ -\infty\ , & \mathbf{b} \in \partial\Omega^d, \end{cases}$$

otherwise it is a smooth function in Ω^d.

Lemma 4.1. *$W_\Omega(\mathbf{b}, H_0)$ has a local minimum inside Ω^d whenever $H_0 \geq H_0(\Omega)$ for some constant $H_0(\Omega)$.*

Proof. Choose a small enough positive constant δ_0 and let

$$\Omega_{\delta_0} = \{x \in \Omega \mid \delta_0 \leq dist(\mathbf{x}, \partial\Omega) \leq 2\delta_0\}\ .$$

If $dist(\mathbf{b}_j, \partial\Omega) \geq \delta_0$ for all j, then $\tilde{g}_\Omega(\mathbf{b}) \geq -M(\delta_0)$, independently of H_0. On the other hand, we can choose d distinct points $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d \in B_R(\mathbf{x}_0)$, where $\mathbf{x}_0$ satisfies $\zeta_1(\mathbf{x}_0) = \min_{\mathbf{x}\in\Omega} \zeta_1(\mathbf{x}) = -m_0 < 0$ such that $B_R(\mathbf{x}_0) \subset \{\mathbf{x} : dist(\mathbf{x}, \partial\Omega) \geq 2\delta_0\}$ and such that

$$\tilde{g}_\Omega(\mathbf{b}) \leq C(R)$$

for some constant $C(R)$ when R is small. Moreover,

$$\sum_{j=1}^d \zeta_1(\mathbf{b}_j) \leq -dm_0 + \tilde{C}_1 R$$

for some constant $\tilde{C}_1$. So,

$$W_\Omega(\mathbf{b}, H_0) \leq -2\pi H_0 d m_0 + C(R) + 2\pi H_0 \tilde{C}_1 R + \frac{1}{2} H_0^2 C(\Omega) \ .$$

If, however, at least one $\mathbf{b}_j$ satisfies $\mathbf{b}_j \in \Omega_{\delta_0}$, then

$$\begin{aligned} W_\Omega(\mathbf{b}, H_0) &\geq -M(\delta_0) - 2\pi H_0 \tilde{C}_2 \delta_0 - 2\pi H_0 (d-1) m_0 + \frac{1}{2} H_0^2 C(\Omega) \\ &\geq -2\pi H_0 d m_0 + C(R) + 2\pi H_0 \tilde{C}_1 R + \frac{1}{2} H_0^2 C(\Omega) \end{aligned}$$

for some constant $\tilde{C}_2$ if R, δ_0 are small enough and H_0 is large enough. Thus, $W_\Omega(\mathbf{b}, H_0)$ has a local minimum inside Ω^d whenever $H_0 > H_0(\Omega)$.

4.2. The existence of vortex solutions. Using the renormalized energy, we now study the solutions of the G–L equations (3)–(6).

THEOREM 4.2. *If $H_0 \geq H_0(\Omega)$ and $W_\Omega(\mathbf{b}, H_0)$ has a nondegenerate local minimum for some $\mathbf{b} \in \Omega^d$, then, for small ϵ, there are solutions $(\psi^\epsilon, \mathbf{A}^\epsilon)$ to the full steady state G–L equations with the gauge choice $\mathbf{A}_\epsilon = \nabla^\perp \zeta_\epsilon$ in Ω, $\mathbf{A}_\epsilon \cdot \mathbf{n} = 0$ on $\partial\Omega$ such that*

$$\psi_\epsilon(\mathbf{x}) \to \psi^*(\mathbf{x}) \quad \text{in } C^{1,\alpha}_{\text{loc}} \left(\overline{\Omega}/\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d\}\right),$$

$$\zeta_\epsilon(\mathbf{x}) \to \zeta^*(\mathbf{x}) \quad \text{in } H^2(\Omega),$$

where

$$-\Delta^2 \zeta^* + \Delta \zeta^* = 2\pi \sum_{j=1}^{d} \delta(\mathbf{b}_j) \quad \text{in } \ \Omega$$

with

$$\zeta^* = 0 \quad \text{on } \partial\Omega,$$

$$\Delta \zeta^* = H_0 \quad \text{on } \partial\Omega,$$

and

$$\psi^*(\mathbf{x}) = \prod_{j=1}^{d} \frac{\mathbf{x} - \mathbf{b}_j}{|\mathbf{x} - \mathbf{b}_j|} e^{ih^*(\mathbf{x})} \quad \text{in } \Omega.$$

Moreover, if we write $\psi^*(\mathbf{x}) = e^{i\theta_b(\mathbf{x}) + ih^*(\mathbf{x})} = e^{i\phi_b}$, then ϕ_b is a multivalued harmonic function on $\Omega \setminus \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d\}$ with $\frac{\partial \phi_b}{\partial n} = 0$ on $\partial\Omega$.

To complete the proof, we use the approach in [21]. We consider the solution of the following system:

$$\frac{\partial \psi}{\partial t} - (\nabla - i\mathbf{A})^2 \psi - \frac{1}{\epsilon^2} \psi (1 - |\psi|^2) = 0 \quad \text{in } \ \Omega, \tag{48}$$

$$\frac{\partial \mathbf{A}}{\partial t} - \Delta \mathbf{A} + \frac{i}{2} (\psi^* \nabla \psi - \psi \nabla \psi^*) + |\psi|^2 \mathbf{A} = 0 \quad \text{in } \Omega, \tag{49}$$

and

$$\operatorname{curl}\mathbf{A} = H_0 \quad \text{on } \partial\Omega, \tag{50}$$

$$(\nabla - i\mathbf{A})\psi\cdot\mathbf{n} = 0 \quad \text{on } \partial\Omega, \tag{51}$$

$$\mathbf{A}\cdot\mathbf{n} = 0 \quad \text{on } \partial\Omega \tag{52}$$

with properly defined initial conditions.

Let

$$\mathcal{E}_\epsilon(\psi,\mathbf{A}) = \frac{1}{4\epsilon^2}(1-|\psi|^2)^2 + \frac{1}{2}\left|(\nabla - i\mathbf{A})\,\psi\right|^2 + \frac{1}{2}|\operatorname{curl}\mathbf{A} - H_0|^2$$

and

$$\tilde{\mathcal{E}}_\epsilon(\psi,\mathbf{A}) = \mathcal{E}_\epsilon(\psi,\mathbf{A}) + \frac{1}{2}|\operatorname{div}\mathbf{A}|^2\ .$$

It is easy to see that (see [11], for example)

$$\int_\Omega \tilde{\mathcal{E}}_\epsilon(\psi(t),\mathbf{A}(t))d\mathbf{x} \le \int_\Omega \tilde{\mathcal{E}}_\epsilon(\psi(0),\mathbf{A}(0))d\mathbf{x}\ .$$

Assuming that the initial condition satisfies

$$\operatorname{div}\mathbf{A}(0) = 0,$$

this in turn implies

$$\int_\Omega \mathcal{E}_\epsilon(\psi(t),\mathbf{A}(t))d\mathbf{x} \le \int_\Omega \mathcal{E}_\epsilon(\psi(0),\mathbf{A}(0))d\mathbf{x}.$$

Let ρ_0 be small enough such that

$$\delta_0 = \min\{dist(\partial\Omega, \partial B_{2\rho_0}(\mathbf{b}_j)), |\mathbf{b}_j - \mathbf{b}_i|, i \ne j,\ i,j = 1,2,\ldots,d\} > 0$$

and

$$\min_{\mathbf{x}\in\partial B_{\rho_0}(\mathbf{b})} W(\mathbf{x}) \ge C(\rho_0) + W(\mathbf{b})\ .$$

Let us construct initial conditions. Let $\mathbf{A}(0) = \nabla^\perp\zeta_b$ in Ω with $\zeta_{\mathbf{b}}$ defined as before. For small ρ, let $B_\rho = \bigcup_{j=1}^d B_\rho(\mathbf{b}_j)$, and let $\psi(0) = e^{i\phi_b}$ in $\Omega\setminus B_\rho$ and extend $\psi(0)$ inside each ball $B_\rho(\mathbf{b}_j)$ by the minimizer of

$$\begin{aligned} I_\rho(\psi,\mathbf{A}) = \min\Bigg\{&\int_{B_\rho(\mathbf{b}_j)}\left(\frac{1}{2}\left|\nabla\psi - i\mathbf{A}\psi\right|^2 + \frac{1}{4\epsilon^2}(1-|\psi|^2)^2\right.\\ &\left.+\frac{1}{2}|\operatorname{curl}\mathbf{A} - H_0|^2\right)d\mathbf{x}\ |\ \ \operatorname{div}\mathbf{A} = 0 \text{ in } \Omega,\quad \mathbf{A}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega,\\ &|\psi|\ |_{\partial B_\rho(\mathbf{b}_j)} = 1\ , (i\psi, (\nabla - i\mathbf{A})\psi\cdot\tau) = g \text{ on } \partial B_\rho(\mathbf{b}_j),\\ &\quad \text{and } deg(\psi, \partial B_\rho(\mathbf{b}_j)) = 1\ , j = 1,2,\ldots,d\Bigg\}, \end{aligned}$$

where τ is the unit tangent vector and $g = \nabla(\theta_b - \zeta_b) \cdot \tau$ on $\partial B_\rho(\mathbf{b}_j)$.

Using the argument in [5] and the gauge invariance, one may show that

$$I_\rho(\psi, \mathbf{A}) = \pi d \log\left(\frac{\rho}{\epsilon}\right) + \gamma d + o(1) + O(\rho).$$

Thus, we may assume

$$E_\epsilon(\psi(0), \mathbf{A}(0)) \leq d\pi \log\frac{1}{\epsilon} + W_\Omega(\mathbf{b}, H_0) + \gamma d + o(1)\ .$$

LEMMA 4.3. *The solution of* (48)–(52) *has the property, for all sufficiently small* ϵ *and for all* $t \geq 0$, *that the set* $\{\mathbf{x} \in \Omega, |\psi_\epsilon(\mathbf{x})| \leq \frac{1}{2}\}$ *is contained in a union of disjoint discs* $B_j, j = 1, 2 \dots, d$, *where*

(i) $B_j = B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$, $\mathbf{x}_j^\epsilon \in B_{\rho_0}(\mathbf{b}_j)$, *for some small* α.

(ii) $\epsilon^\alpha \int_{\partial B_j} \mathcal{E}_\epsilon(\psi_\epsilon, \mathbf{A}_\epsilon) d\mathbf{x} \leq C(K_0)$.

Thus, $d_j = deg(\partial B_j, \psi)$ *is well defined and* $d_j = 1, j = 1, 2, \dots, d$.

Proof. Let us define

$$P = \{t \geq 0 \ \mid \ (\psi_\epsilon, \mathbf{A}_\epsilon) \text{ s.t. (i)} - \text{(ii)}\}\ .$$

Clearly, P is a closed set and by choosing the initial conditions properly, we have $0 \in P$. We now show that P is open; thus $P = [0, \infty)$. First, we have the following claim.

Claim. For any $t \in P$, there exists a $\rho^* \in (0, \rho_0)$, independent of sufficiently small ϵ, such that $\min\{dist(\mathbf{x}^\epsilon, \partial B_{\rho_0}(\mathbf{b})),\} \geq \rho^*$.

Assume that the claim is true. Then for any $t' > t$ but sufficiently close to t,

$$deg(\partial B_{\rho^*}(\mathbf{x}_j^\epsilon), \psi_\epsilon(\cdot, t')) = 1.$$

We may then follow the ideas in [21] and [5] to construct, for $\psi_\epsilon(\cdot, t')$, a new disc $\tilde{B}_j$ which satisfies (i)–(ii). Again, the center of the disc can be shown to satisfy the above claim. Thus, the lemma is proved.

Now we verify the claim. Suppose the claim is not true, there exists a sequence $\epsilon_n \searrow 0$, $t_n \nearrow t^* \in P$, such that $\mathbf{x}_i^{\epsilon_n} \to \bar{\mathbf{b}}_i \in B_{\rho_0}(\mathbf{b}_i)$ for $i = 1, 2 \dots, d$, and $\bar{\mathbf{b}}_j \in \partial B_{\rho_0}(\mathbf{b}_j)$ for some j.

We will follow constructions similar to those in [21] and [5] to show that one may replace $(\psi_\epsilon, \mathbf{A}_\epsilon)$ by $(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)$ to satisfy

$$E(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon) \geq \pi d \log\frac{1}{\epsilon} + W_\Omega(\bar{\mathbf{b}}, H_0) + \gamma d + o(1).$$

First, let us do gauge transformation such that $\bar{\mathbf{A}}_\epsilon = \nabla^\perp \zeta_\epsilon$ with $\zeta_\epsilon = 0$ on $\partial\Omega$. Let us consider

$$\min\left\{ \int_{\Omega \setminus \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\bar{\psi}, \bar{\mathbf{A}}) d\mathbf{x} \ \mid \ \bar{\mathbf{A}} = \nabla^\perp \zeta \text{ in } \Omega\ ,\ deg(\bar{\psi}, \partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)) = 1, \right.$$

$$\left. |\bar{\psi}||_{\partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} = |\psi_\epsilon||_{\partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)}\ , \text{ and } (i\bar{\psi}, (\nabla - i\bar{\mathbf{A}})\bar{\psi} \cdot \tau) = g_\epsilon \ \text{ on } \partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon) \right\},$$

where $g_\epsilon = (i\psi_\epsilon, (\nabla - i\mathbf{A}_\epsilon)\psi_\epsilon \cdot \tau)$. Note that the boundary conditions are the same gauge invariant boundary conditions considered in [5] and α satisfies property (ii) listed above.

Let $(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)$ denote a minimizer. Let $\bar{\psi}_\epsilon = \bar{\rho}_\epsilon e^{i\bar{\phi}_\epsilon}$ and $\psi_\epsilon = \rho_\epsilon e^{i\phi_\epsilon}$ on the boundary of $\bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$. To extend it inside the ϵ^α balls, we define a gauge transformation by

$$\Delta^2 \chi = 0 \quad \text{in } \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$$

with boundary conditions

$$\kappa\chi = \bar{\phi}_\epsilon - \phi_\epsilon \quad \text{ and } \quad \frac{\partial \chi}{\partial \mathbf{n}} = 0 \qquad \text{on } \bigcup_{j=1}^d \partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon) \ .$$

Note that χ is well defined on $\bigcup_{j=1}^d \partial B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$. Thus, inside $\bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$, we define

$$\bar{\psi}_\epsilon = \psi_\epsilon e^{i\kappa\chi} \quad \text{ and } \quad \bar{\mathbf{A}}_\epsilon = \mathbf{A}_\epsilon + \nabla\chi \ .$$

By the gauge invariance and the choice of the small α, we have the energy lower bound inside the ϵ^α balls:

$$\begin{aligned} \int_{\bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon) d\mathbf{x} &= \int_{\bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\psi_\epsilon, \mathbf{A}_\epsilon) d\mathbf{x} \\ &\geq \pi d \log\left(\frac{\epsilon^\alpha}{\epsilon}\right) - C \end{aligned}$$

for small enough ϵ. So, we have the energy upper bound outside:

$$\int_{\Omega \setminus \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon) d\mathbf{x} \leq \pi d\alpha \log\left(\frac{1}{\epsilon}\right) + C \ .$$

Using arguments similar to those in [21] and [5], we have

$$|\bar{\psi}_\epsilon| \geq \frac{1}{2} \quad \text{in } \Omega \setminus \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon).$$

Moreover, we may modify the arguments of [21] and [5] to show that we have strong convergence of

$$\bar{\psi}_\epsilon \to e^{i\phi_{\bar{\mathbf{b}}}}$$

and

$$\bar{\mathbf{A}}_\epsilon \to \nabla^\perp \zeta_{\bar{\mathbf{b}}}$$

outside any small neighborhood of $\bar{\mathbf{b}}_j, j = 1, 2, \ldots, d$. Indeed, using the fact that $H = \operatorname{curl} \bar{\mathbf{A}}_\epsilon$ satisfies the equation

$$\operatorname{div}\left(\frac{1}{|\bar{\psi}_\epsilon|^2} \nabla H\right) = H \quad \text{in } \Omega \setminus \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon) \ ,$$

we note that $\rho = |\bar{\psi}_\epsilon| > 1/2$ in $\Omega \setminus \bigcup_{j=1}^d B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)$ (in fact, ρ is arbitrarily close to 1 if we allow ϵ small). Therefore, we obtain from elliptic estimates that the $W^{1,p}$ norm and the C^γ norm of H are bounded locally for some $p > 2$ and $\gamma > 0$. On the other hand, from Lemma 4.1 in [21] and arguments in [21] and [5], one deduces the local strong convergence of $\bar{\psi}_\epsilon$. Then the assertion on the convergence of $\bar{\mathbf{A}}_\epsilon$ follows. We omit the details.

Next, for small δ, by strong convergence,

$$\begin{aligned} &\int_{\Omega\setminus\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)d\mathbf{x} \\ &\qquad \geq \pi d \log\left(\frac{1}{\delta}\right) + W_\Omega(\bar{\mathbf{b}}, H_0) + o(1) + O(\delta). \end{aligned} \tag{53}$$

We can also get $deg(\bar{\psi}_\epsilon, \partial B_\delta(\bar{\mathbf{b}}_j)) = 1$ for any j.

Let us write $\bar{\psi}_\epsilon = |\bar{\psi}_\epsilon| e^{i\phi_\epsilon} = \rho e^{i\theta + ih}$, where θ is the angle function inside each $B_\delta(\mathbf{x}_j^\epsilon)$ so that

$$e^{i\theta(\mathbf{x})} = \frac{\mathbf{x} - \bar{\mathbf{b}}_j}{|\mathbf{x} - \bar{\mathbf{b}}_j|} \quad \text{in } B_\delta(\mathbf{x}_j^\epsilon)\ \forall\, j.$$

We have, by gauge invariance, the energy upper bound outside the ϵ^α balls:

$$\int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \left(|\nabla\rho|^2 + \rho^2|\nabla\theta - \nabla^\perp\zeta|^2 + |\Delta\zeta|^2\right) d\mathbf{x} \ \leq C\log\left(\frac{1}{\epsilon}\right) + K, \tag{54}$$

where $\nabla^\perp\zeta = \bar{\mathbf{A}}_\epsilon$ is the magnetic potential after a gauge transformation.

Thus, we have

$$\begin{aligned} &\int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} (1-\rho^2)|\nabla\theta - \nabla^\perp\zeta|^2 d\mathbf{x} \\ &\leq \int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \left[2(1-\rho^2)|\nabla^\perp\zeta|^2 + 2(1-\rho^2)|\nabla\theta|^2\right] d\mathbf{x} \\ &\leq \int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \left(\frac{1}{\epsilon}(1-\rho^2)^2 d\mathbf{x} + \epsilon|\nabla^\perp\zeta|^4\right) d\mathbf{x} + o(1) \\ &\leq C\epsilon\log\left(\frac{1}{\epsilon}\right) + \epsilon\left(\int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} |\Delta\zeta|^2 d\mathbf{x}\right)^2 + o(1) \\ &\leq C\epsilon\log\left(\frac{1}{\epsilon}\right) + C\epsilon\left(\log\left(\frac{1}{\epsilon}\right)\right)^2 + o(1) \\ &\leq o(1), \end{aligned}$$

where C is some generic constant.

This implies in particular that

$$\begin{aligned} &\int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} \mathcal{E}(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon) d\mathbf{x} \\ &\quad \geq \int_{\bigcup_{j=1}^d B_\delta(\mathbf{x}_j^\epsilon)\setminus B_{\epsilon^\alpha}(\mathbf{x}_j^\epsilon)} |\nabla\theta - \nabla^\perp\zeta|^2 d\mathbf{x} + o(1) \\ &\quad \geq \pi d\log\left(\frac{\delta}{\epsilon^\alpha}\right) + o(1). \end{aligned}$$

Therefore, by arguments in [5] and the convergence of $(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)$, we have

$$\sum_{j=1}^{d} \int_{B_\delta(\mathbf{x}_j^\epsilon)} \mathcal{E}_\epsilon(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)\, d\mathbf{x} \geq \pi d \log\left(\frac{\delta}{\epsilon}\right) + \gamma d + o(1) + O(\delta). \tag{55}$$

Thus, for small ϵ and small δ,

$$\int_\Omega \mathcal{E}_\epsilon(\bar{\psi}_\epsilon, \bar{\mathbf{A}}_\epsilon)\, d\mathbf{x} \geq \pi d \log\left(\frac{1}{\epsilon}\right) + W_\Omega(\bar{\mathbf{b}}, H_0) + \gamma d + o(1)\ .$$

This leads to a contradiction of the energy upper bound obtained from the energy dissipation from the energy of the initial condition since $\mathbf{b}$ is the nondegenerate local minimum of $W(\mathbf{b}, H_0)$. Hence, the claim is true.

To complete the proof of the theorem, let us use the uniform bound on the solution $(\psi^\epsilon, \mathbf{A}^\epsilon)$ as $t \to \infty$. We get a subsequence t_n such that $(\psi^\epsilon(t_n), \mathbf{A}^\epsilon(t_n)) \to (\psi^\epsilon, \mathbf{A}^\epsilon)$ as $t \to \infty$. One may then easily check that $(\psi^\epsilon, \mathbf{A}^\epsilon)$ is the critical point of the G–L functional. This proves the theorem.

Remark.

(1) One may prove, using arguments given in [5], that as $\epsilon \to 0$, the zeros of the ψ^ϵ go to the local minimum of the renormalized energy.

(2) By a more careful analysis, one may replace the assumption that the renormalized energy has a nondegenerate local minimum by simply the existence of a local minimum.

(3) One can also prove, via [19], the existence of general critical points (saddle points) of the G–L functional under similar conditions.

(4) According to the nondimensionalization used here, our theorem describes the phenomenon that the G–L system has a vortex solution for an applied field of strength on the order of $\frac{1}{\kappa}$. Recall that the standard estimates for the lower critical field are $\frac{\log(\kappa)}{\kappa}$. That is, we can prove the existence of a stable vortex state below H_{c1}.

4.3. The weak hysteresis near the lower critical field $\boldsymbol{H_{c1}}$ for type-II superconductors. We have just shown that when we decrease the applied magnetic field, there may be stable vortex states (even local energy minimizing states) even when the field strength is below the lower critical field $H_{c1}(\approx \log\kappa/\kappa)$, say, c_0/κ for some constant c_0. On the other hand, it is rather straightforward to check from the definition of $W_\Omega(\mathbf{b}, H_0)$ that there is no local minimum of the function $W_\Omega(\mathbf{b}, H_0)$ when H_0 is sufficiently small. That is, we have shown the existence of a subcooling field $0 < H_{sc} < H_{c1}$.

Let us consider another case where we gradually increase the applied field. We start, say, with a perfect superconducting state (or the *Meissner* state). It is quite possible that the Meissner state exists even when the applied field is much larger than the lower critical field H_{c1}. Indeed, if one looks for a solution to the full steady state G–L equations $(\psi, \mathbf{A})$ with $|\psi| \neq 0$, then one may write $\psi = fe^{i\chi}$ for some $f \neq 0$.

Using gauge invariance, we define $\mathbf{Q} = \mathbf{A} - \frac{1}{\kappa}\nabla\chi$, then $(f, \mathbf{Q})$ remains a solution of the steady state G–L equations which can now be written in the following form:

$$\left\{ \begin{array}{l} \frac{1}{\kappa^2}\Delta f = f^3 - f + f|\mathbf{Q}|^2 \\ -\mathrm{curl}^2\mathbf{Q} = f^2\mathbf{Q} \end{array} \right. \quad \text{in } \Omega \tag{56}$$

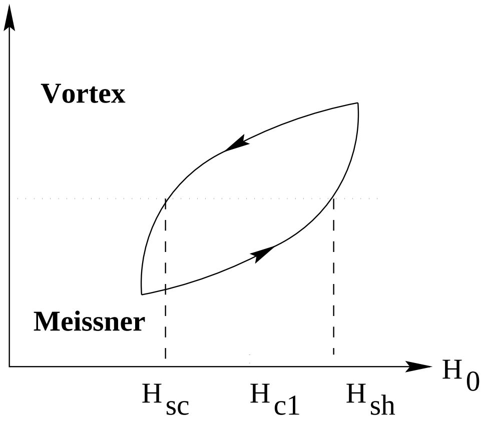


FIG. 8. *The weak hysteresis diagram for type-*II *superconductors near* H_{c1}.

with boundary conditions

$$\begin{cases} \frac{\partial f}{\partial n} = 0 \\ \mathbf{Q} \cdot \mathbf{n} = 0 \\ H = \mathrm{curl}\mathbf{Q} = H_0 \end{cases} \quad \text{on } \partial\Omega. \tag{57}$$

From the equation for $\mathbf{Q}$, we obtain

$$\begin{cases} - \quad \frac{\partial H}{\partial x_2} = f^2 Q_1 \\ \qquad \frac{\partial H}{\partial x_1} = f^2 Q_2 \end{cases} \quad \text{in } \Omega,$$

or, equivalently,

$$\mathrm{div}\left(\frac{\nabla H}{f^2}\right) = H \quad \text{in } \Omega.$$

As in [2], we let $\kappa \to \infty$ to obtain

$$f = 1 - |\mathbf{Q}|^2.$$

So, $|\nabla H|^2 = f^4(1 - f^2)$. Let $f = \rho(|\nabla H|)$. Then the equation for H becomes

$$\begin{cases} \mathrm{div}\,(\rho(\nabla H)\nabla H) = H \quad \text{in } \Omega, \\ H = \ H_0 \quad \text{on } \partial\Omega\ . \end{cases} \tag{58}$$

For $H_0 > 0$ small enough (but obviously larger than $\lim_{\kappa\to\infty} \frac{\log\kappa}{\kappa} = 0$), one can show the existence of a solution to the above equation. Moreover, such a solution is a linearly stable wholly superconducting state (Meissner state) for $0 < H_0 \le H_0^*$. The new field strength H_0^* is called the superheating field; see [2] for details.

Combining this latter result with ours given in section 4.2, a weak-hysteresis diagram of type-II superconductors near the lower critical field H_{c1} is completed (see Figure 8).

REFERENCES

[1] N. Andre and I. Shafrir, *Asymptotic Behavior of Minimizers for the Ginzburg–Landau Functional with Weights,* I *and* II, preprint.

[2] H. Berestycki, A. Bonnet, and S. Chapman, *A semi-elliptic system arising in the theory of type*-II *superconductivity*, Comm. Appl. Nonlinear Anal., 1 (1994), pp. 1–21.

[3] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg–Landau Vortices*, Birkhäuser Boston, Cambridge, MA, 1994.

[4] P. Bauman, D. Phillips, and Q. Tang, *Stable Nucleation for the Ginzburg–Landau System with an Applied Magnetic Field*, preprint.

[5] F. Bethuel and T. Riviere, *Vortices for a Variational Problem Related to Superconductivity*, preprint.

[6] S. Chapman, Q. Du, and M. Gunzburger, *A variable thickness model for superconducting thin films*, Z. Angew. Math. Phys., 47 (1995), pp. 410–431.

[7] Z. Chen, K. H. Hoffmann, and L. S. Jiang, *On the Lawrence-Doniach model for layered superconductors*, European J. Appl. Math, to appear.

[8] Z. Chen, C. M. Elliot, and Q. Tang, *Justification of a Two-Dimensional Evolutionary Ginzburg–Landau Superconductivity Model*, preprint.

[9] S. Ding and Z. Liu, *Pinning of Vortices for a Variational Problem Related to the Superconducting Thin Film having Variable Thickness*, preprint.

[10] P. De Gennes, *Superconductivity of Metals and Alloys*, Addison-Wesley Publishing Company, Reading, MA.

[11] Q. Du, *Global existence and uniqueness of solutions of the time-dependent Ginzburg–Landau model for superconductivity*, Appl. Anal., 53 (1994), pp. 1–17.

[12] Q. Du and M. D. Gunzburger, *A model for super-conducting thin films having variable thickness*, Phys. D, 69 (1993), pp. 215–231.

[13] Q. Du, M. D. Gunzburger, and J. Peterson, *Analysis and approximation of Ginzburg–Landau models for superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[14] Q. Du, M. D. Gunzburger, and J. Peterson, *Computational simulations of type*-II *superconductivity including pinning mechanisms*, Phys. Rev. B. (3), (1995), pp. 16194–16203.

[15] W. E, *Dynamics of vortices in Ginzburg–Landau theories and applications to superconductivity*, Phys. D, 77 (1994), pp. 383–404.

[16] W. E, *Dynamics of vortex liquid in Ginzburg–Landau theories, with applications to superconductivity*, Phys. Rev. B. 50, (1994), pp. 1126–1135.

[17] L. Gorkov and G. Eliashberg, *Generalization of the Ginzburg–Landau equations for nonstationary problems in the case of alloys with paramagnetic impurities*, Soviet Phys. JETP, 27 (1968), pp. 328–334.

[18] C. Lefter and V. D. Radulescu, *On the Ginzburg–Landau energy with weight*, C. R. Acad. Sci. Paris, to appear.

[19] C. Lin and F. H. Lin, *Minimax solutions of Ginzburg–Landau equations*, Select Math. (N.S.), 3 (1991), pp. 99–113.

[20] F. H. Lin, *Solutions of Ginzburg–Landau equations and critical points of the renormalized energy*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 599–622.

[21] F. H. Lin, *Some dynamic properties of Ginzburg–Landau vortices*, Comm. Pure Appl. Math., 49 (1996), pp. 323–364.

[22] F. H. Lin, *Mixed vortex and anti-vortex solutions of Ginzburg–Landau equations*, Arch. Rational Mech. Anal., 133 (1995), pp. 103–127.

[23] J. Neu, *Vortices in complex scalar fields*, Phys. D, 43 (1990), pp. 385–406.

[24] L. Pisman and J. D. Rodriguez, *Mobilities of singularities in dissipative Ginzburg–Landau equations*, Phys. Rev. A, 42 (1990), pp. 2471–2474.

[25] J. Rubinstein and P. Sternberg, *On the Slow Motion of Vortices in the Ginzburg–Landau Heat Flow*, preprint.

[26] L. Simon, *Lectures on geometric measure theory*, Austral. Nat. Univ. CMA (1983).

[27] L. Simon, *Asymptotics for a class of non-linear evolution equations with applications to geometric problems*, Ann. of Math, 118 (1983), pp. 525–572.

[28] Q. Tang and S. Wang, *Time dependent Ginzburg–Landau equations of superconductivity*, Physica D, 88 (1995), pp. 139–166.

SIAM J. Math. Anal.
Vol. 28, No. 6, pp. 1294–1308, November 1997

SPACE HOMOGENEOUS SOLUTIONS TO THE CAUCHY PROBLEM FOR SEMICONDUCTOR BOLTZMANN EQUATIONS[*]

A. MAJORANA[†] AND S. A. MARANO[†]

Abstract. The nonlinear Boltzmann equation for an electron gas in a semiconductor is investigated. Some meaningful properties of the collision operator are first presented. A large class of kernels is allowed. Then the global existence and uniqueness of bounded, continuous, space-independent solutions to the related Cauchy problem is performed. Finally, the conservation of mass is examined.

Key words. Boltzmann equation, semiconductor, Dirac distribution, bounded continuous solution, existence and uniqueness

AMS subject classifications. 45K05, 46F10, 82C40

PII. S0036141095291397

1. Introduction.

A well-accepted model for the charge carrier transport in semiconductors is the Boltzmann equation for an unknown function defined in the seven-dimensional phase space spanned by time, space coordinates, and wave vector [17], [20]. This equation contains a very complicated nonlinear integral operator (the collision term). So, to avoid technical difficulties, numerous simplifying assumptions and modifications are often introduced. Although such *approximate* new Boltzmann-like models provide acceptable results in a few cases, properties of the true equation are frequently lost [11], [12], [16]. Therefore, precise information may be obtained only through the exact equation. For example, the increasing miniaturization of modern electron devices requires this accurate modeling in order to correctly describe the evolution of physical parameters. An alternative approach, which avoids the use of the Boltzmann equation, is furnished by the direct molecular simulation [10]. In this framework, techniques like Monte Carlo simulation provide detailed information about carrier transport within advanced devices, but the computational burden limits its use for many devices' engineering applications. For such reasons, great attention has recently been paid to the study of the semiconductor Boltzmann equation.

In this paper we consider the Boltzmann equation for the distribution function f of an electron gas in a semiconductor with the full collision operator $Q(f)$, which describes the interactions between the electron gas and the molecules, assumed in thermal equilibrium, of the semiconductor. We omit electron–electron collisions and hole–electron interactions, because they often are physically negligible with respect to the other interactions. This avoids both cumbersome formulas for $Q(f)$ and technical troubles that otherwise might occur (see section 4).

The main difficulty in studying the equation arises from the fact that the collision term contains integral operators of the type

$$I(f)(t,\mathsf{x},\mathsf{k}) = \int_{\mathbb{R}^3} \mathcal{G}(\mathsf{k},\mathsf{k}') f(t,\mathsf{x},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}', \quad (t,\mathsf{x},\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3,$$

[*] Received by the editors September 8, 1995; accepted for publication (in revised form) June 26, 1996.

http://www.siam.org/journals/sima/28-6/29139.html

[†] Department of Mathematics, University of Catania, Viale A. Doria 6, 95125 Catania, Italy (majorana@dipmat.unict.it, marano@dipmat.unict.it).

where $\mathcal{G}$ denotes a continuous function, ε represents the particle energy, μ is a constant, and $\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)$ means the composition of the real-valued function $(\mathsf{k}, \mathsf{k}') \to \varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu$ and the Dirac distribution δ.

Usually, the distribution δ is replaced by a smooth function (we refer, for instance, to [14], [15], [18], [19]), or a simple expression of ε is employed (see, for example, [11], [16]) so that $I(f)$ may be reduced to an integral operator without distributions. However, trustable numerical results for the high-energy part of the electron distribution function can be achieved only when the correct form of ε is used [21].

We make quite general assumptions on ε, which cover all of the most common expressions considered in applications and simulations. Under such hypotheses, in sections 1 and 2, we show that the collision operator is well defined and has significant properties in the space of all bounded continuous functions. This is performed by first giving a precise meaning for and then proving some basic facts about the integral operator $I(f)$. In section 3, we examine the Boltzmann equation without external fields, since f does not depend on space coordinates. We study the related Cauchy problem and we establish the global existence and uniqueness of bounded continuous solutions. Further, we show that if the initial datum is integrable on $\mathbb{R}^3$, then the same holds for the function $\mathsf{k} \to f(t, \mathsf{k})$, $t \in \mathbb{R}_0^+$, and its integral is constant with respect to the time t. Throughout the paper, *integrable* always means Lebesgue integrable.

2. Basic equations.

The Boltzmann equation for an electron gas reads

$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_{\mathsf{k}}\varepsilon \cdot \nabla_{\mathsf{x}} f - \frac{e}{\hbar}\mathsf{E} \cdot \nabla_{\mathsf{k}} f = Q(f), \tag{2.1}$$

where the unknown function $(t, \mathsf{x}, \mathsf{k}) \to f(t, \mathsf{x}, \mathsf{k})$ represents the existence probability of an electron at the position $\mathsf{x} \in \mathbb{R}^3$, with the wave vector $\mathsf{k} \in \mathbb{R}^3$ at time $t \in \mathbb{R}_0^+$ [13], [17]. The parameters $\hbar$ and e are the Planck constant divided by 2π and the positive electric charge, respectively. The symbol ∇_{k} stands for the gradient with respect to the variables k and ∇_{x} with respect to the space coordinates x. The particle energy ε is an assigned nonnegative function defined in $\mathbb{R}^3$. In (2.1) the external force represents the electric field $(t, \mathsf{x}) \to \mathsf{E}(t, \mathsf{x})$, which satisfies a suitable Poisson equation.

We follow a semiclassical approach for the collision term $Q(f)$, so

$$Q(f) = \int_{\mathbb{R}^3} \left[S(\mathsf{k}', \mathsf{k}) f'(1 - f) - S(\mathsf{k}, \mathsf{k}') f(1 - f')\right] d\mathsf{k}' . \tag{2.2}$$

Here, for any function ϕ we use the notation $\phi' = \phi(\mathsf{k}')$. The kernel S is defined by

$$S(\mathsf{k}, \mathsf{k}') = \sum_{i=1}^{n} \mathcal{G}_i(\mathsf{k}, \mathsf{k}') \left[(n_i + 1)\delta(\varepsilon' - \varepsilon + \hbar\omega_i) + n_i \delta(\varepsilon' - \varepsilon - \hbar\omega_i)\right], \tag{2.3}$$

where n is a fixed positive integer, $\mathcal{G}_i : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_0^+$ denotes a continuous function satisfying $\mathcal{G}_i(\mathsf{k}, \mathsf{k}') = \mathcal{G}_i(\mathsf{k}', \mathsf{k})$ for every k, $\mathsf{k}' \in \mathbb{R}^3$, and n_i and ω_i are nonnegative constants, $i = 1, 2, ..., n$. The symbol $\delta(\varepsilon' - \varepsilon \pm \hbar\omega_i)$ means the composition of the real-valued function $(\mathsf{k}, \mathsf{k}') \to \varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) \pm \hbar\omega_i$ and the Dirac distribution δ (see [8, Chapter III]).

The kernel (2.3) is rather different from the one appearing in the widely known Boltzmann equation for a perfect rarefied gas [7]. Indeed, in this case, the Dirac distributions have been ab initium eliminated in some way (we refer to [4] and [5] for a complete treatment). Consequently, the kernel of the integral term is a usual measurable function, and (2.1) can be solved in several settings, for example, Lebesgue spaces. This is not the present case. In fact, since the function ε may have many different expressions, it is not possible in general, by using a unique technique, to transform the integral operator (2.2) to an equivalent one without Dirac distributions. Moreover, the presence of distributions requires the use of functions which are at least continuous with respect to the wave vector.

We consider here the Boltzmann equation (2.1) in the simpler case when $\mathsf{E} \equiv 0$. Further, we suppose that the unknown function $(t, \mathsf{x}, \mathsf{k}) \to f(t, \mathsf{x}, \mathsf{k})$ does not depend on $\mathsf{x} \in \mathbb{R}^3$; namely, it is space homogeneous. This allows us to treat the equation with the *original expression* (2.3) *for the kernel* S.

To the best of our knowledge, no existence results concerning the complete unmodified equation (2.1) have been established. Some authors [14], [15], [19] studied the Cauchy problem for (2.1) with the full differential operator but *a different kernel*, which is usually taken to be smooth and without Dirac distributions. This choice seems to be forced by the used techniques, which require the existence of partial derivatives of S. Of course, such assumption is not satisfied by kernel (2.3). So, a new approach is probably necessary in order to get strong solutions of the full equation (2.1).

3. The collision operator. We will now make the following assumptions on the function ε:

(a$_1$) $\varepsilon : \mathbb{R}^3 \to \mathbb{R}_0^+$ *is continuous.*

(a$_2$) *There exists* $\mathsf{k}_0 \in \mathbb{R}^3$ *such that if* $D = [0, 2\pi] \times [0, \pi]$,

$$\mathsf{n} = (\sin\varphi\cos\vartheta, \sin\varphi\sin\vartheta, \cos\varphi), \quad \textit{and} \quad \eta(\rho, \vartheta, \varphi) = \varepsilon(\mathsf{k}_0 + \rho\mathsf{n}),$$

$\rho \geq 0$, $(\vartheta, \varphi) \in D$, *then*

(a$_{21}$) *the function* η *admits a continuous and positive partial derivative with respect to* ρ *for every* $(\rho, \vartheta, \varphi) \in]0, +\infty[\times D$,

(a$_{22}$) $\lim_{\rho \to 0^+} \rho^2 \left[\frac{\partial \eta(\rho, \vartheta, \varphi)}{\partial \rho} \right]^{-1} = 0$ *uniformly in* $(\vartheta, \varphi) \in D$,

(a$_{23}$) $\lim_{\rho \to +\infty} \eta(\rho, \vartheta, \varphi) = +\infty$ *uniformly in* $(\vartheta, \varphi) \in D$.

In Appendix A we list the most common expressions of ε considered in applications and simulations [3], [10]. It is easily seen that each of them satisfies hypotheses (a$_1$) and (a$_2$).

For every $f \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$, we define

$$A(f)(t, \mathsf{x}, \mathsf{k}) = \int_{\mathbb{R}^3} S(\mathsf{k}', \mathsf{k}) f(t, \mathsf{x}, \mathsf{k}')\, d\mathsf{k}', \tag{3.1}$$

$$A^*(f)(t, \mathsf{x}, \mathsf{k}) = \int_{\mathbb{R}^3} S(\mathsf{k}, \mathsf{k}') f(t, \mathsf{x}, \mathsf{k}')\, d\mathsf{k}', \tag{3.2}$$

$$B(f)(t, \mathsf{x}, \mathsf{k}) = \int_{\mathbb{R}^3} S(\mathsf{k}, \mathsf{k}') \left[1 - f(t, \mathsf{x}, \mathsf{k}')\right] d\mathsf{k}', \tag{3.3}$$

$$\nu(\mathsf{k}) = \int_{\mathbb{R}^3} S(\mathsf{k}, \mathsf{k}')\, d\mathsf{k}',$$

$(t, \mathsf{x}, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3$, where the kernel S is given by (2.3). Each of the above expressions is a finite sum of terms of the kind

$$\Phi(t, \mathsf{x}, \mathsf{k}) = \int_{\mathbb{R}^3} \phi(t, \mathsf{x}, \mathsf{k}, \mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}', \tag{3.4}$$

where ϕ indicates a function belonging to $C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3)$ and μ is a real number. In Appendix B we first give a precise meaning to the right-hand side of (3.4) and then prove some technical lemmas which guarantee the following properties for Φ:

(R$_1$) $\Phi \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$.

(R$_2$) *For every $\psi \in C^0(\mathbb{R}^3)$ with compact support,*

$$\int_{\mathbb{R}^3} \psi(\mathsf{k})\Phi(t,\mathsf{x},\mathsf{k})\,d\mathsf{k} = \int_{\mathbb{R}^3} d\mathsf{k}' \int_{\mathbb{R}^3} \psi(\mathsf{k})\phi(t,\mathsf{x},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\,. \tag{3.5}$$

(R$_3$) *If $\phi(t,\mathsf{x},\mathsf{k},\mathsf{k}') \geq 0$ in $\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3$ then the function $\mathsf{k} \to \Phi(t,\mathsf{x},\mathsf{k})$ is integrable on $\mathbb{R}^3$ for any $(t,\mathsf{x}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$ and* (3.5) *holds for all nonnegative bounded $\psi \in C^0(\mathbb{R}^3)$.*

The aim of this section is to point out some meaningful properties of the operators $A(f)$, $B(f)$, and $Q(f)$. Throughout the paper, we suppose

$$M_A = \sup\left\{A(1)(\mathsf{k}) : \mathsf{k} \in \mathbb{R}^3\right\} < +\infty,\; M_B = \sup\left\{B(0)(\mathsf{k}) : \mathsf{k} \in \mathbb{R}^3\right\} < +\infty. \tag{3.6}$$

Since one has $B(f) = \nu - A^*(f)$, due to (2.2), we can write

$$Q(f) = (1-f)A(f) - fB(f). \tag{3.7}$$

As a consequence of (R$_1$), we immediately infer the following proposition.

PROPOSITION 3.1. *Let $f \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$. Then $A(f)$, $B(f) \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$. Consequently, $Q(f) \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$. If, moreover, $0 \leq f(t,\mathsf{x},\mathsf{k}) \leq 1$ for every $(t,\mathsf{x},\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3$, then*

$$0 \leq A(f)(t,\mathsf{x},\mathsf{k}) \leq M_A \quad \textit{and} \quad 0 \leq B(f)(t,\mathsf{x},\mathsf{k}) \leq M_B \quad \textit{in } \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3.$$

We now come to the integrability of the function $\mathsf{k} \to Q(f)(t,\mathsf{x},\mathsf{k})$.

PROPOSITION 3.2. *Let $f \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3)$ be such that the function $\mathsf{k} \to f(t,\mathsf{x},\mathsf{k})$ is integrable on $\mathbb{R}^3$ for all $(t,\mathsf{x}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$ and $0 \leq f(t,\mathsf{x},\mathsf{k}) \leq 1$ in $\mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3$. Then the functions $\mathsf{k} \to A(f)(t,\mathsf{x},\mathsf{k})$ and $\mathsf{k} \to Q(f)(t,\mathsf{x},\mathsf{k})$ are integrable on $\mathbb{R}^3$ for every $(t,\mathsf{x}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$. Furthermore, $\int_{\mathbb{R}^3} Q(f)(t,\mathsf{x},\mathsf{k})\,d\mathsf{k} = 0$.*

Proof. Fix $(t,\mathsf{x}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$. Let $\psi_n : \mathbb{R} \to \mathbb{R}_0^+$, $n \in \mathbb{N}$, be defined by setting, for $z \in \mathbb{R}$,

$$\psi_n(z) = \begin{cases} 1 & \text{if} \quad |z| \leq n, \\ 1+n-|z| & \text{if} \quad n < |z| < n+1, \\ 0 & \text{if} \quad |z| \geq n+1. \end{cases} \tag{3.8}$$

Owing to (3.5) and Proposition 3.1, for any $n \in \mathbb{N}$ one has

$$\begin{aligned} \int_{\mathbb{R}^3} A(f)(t,\mathsf{x},\mathsf{k})\psi_n(|\mathsf{k}|)\,d\mathsf{k} &= \int_{\mathbb{R}^3} d\mathsf{k}\,\psi_n(|\mathsf{k}|) \int_{\mathbb{R}^3} S(\mathsf{k}',\mathsf{k})f(t,\mathsf{x},\mathsf{k}')\,d\mathsf{k}' \\ &= \int_{\mathbb{R}^3} d\mathsf{k}'\,f(t,\mathsf{x},\mathsf{k}') \int_{\mathbb{R}^3} S(\mathsf{k}',\mathsf{k})\psi_n(|\mathsf{k}|)\,d\mathsf{k} \leq M_A \int_{\mathbb{R}^3} f(t,\mathsf{x},\mathsf{k}')\,d\mathsf{k}'\,. \end{aligned}$$

Since $A(f)(t,\mathsf{x},\mathsf{k})\psi_n(|\mathsf{k}|) \leq A(f)(t,\mathsf{x},\mathsf{k})\psi_{n+1}(|\mathsf{k}|)$ and

$$\lim_{n\to+\infty} A(f)(t,\mathsf{x},\mathsf{k})\psi_n(|\mathsf{k}|) = A(f)(t,\mathsf{x},\mathsf{k})$$

pointwise in $\mathbb{R}^3$, the monotone convergence theorem ensures that the function $\mathsf{k} \to A(f)(t, \mathsf{x}, \mathsf{k})$ is integrable on $\mathbb{R}^3$. Moreover,

$$\int_{\mathbb{R}^3} A(f)(t,\mathsf{x},\mathsf{k})\, d\mathsf{k} \le M_A \int_{\mathbb{R}^3} f(t,\mathsf{x},\mathsf{k})\, d\mathsf{k}\,. \tag{3.9}$$

The integrability of $\mathsf{k} \to Q(f)(t,\mathsf{x},\mathsf{k})$ is an immediate consequence of (3.7) because, by Proposition 3.1, the function $B(f)$ is bounded.

Now, let $\psi \in C^0(\mathbb{R}^3)$ be nonnegative and bounded. Bearing in mind (3.7) and (R_3), we see that

$$\begin{aligned}\int_{\mathbb{R}^3} Q(f)(t,\mathsf{x},\mathsf{k})\psi(\mathsf{k})\, d\mathsf{k} &= \int_{\mathbb{R}^3} d\mathsf{k}\,(1-f(t,\mathsf{x},\mathsf{k}))\psi(\mathsf{k}) \int_{\mathbb{R}^3} S(\mathsf{k}',\mathsf{k}) f(t,\mathsf{x},\mathsf{k}')\, d\mathsf{k}' \\ &\quad - \int_{\mathbb{R}^3} d\mathsf{k}\, f(t,\mathsf{x},\mathsf{k})\psi(\mathsf{k}) \int_{\mathbb{R}^3} S(\mathsf{k},\mathsf{k}')(1-f(t,\mathsf{x},\mathsf{k}'))\, d\mathsf{k}' \\ &= \int_{\mathbb{R}^3} d\mathsf{k}'\, f(t,\mathsf{x},\mathsf{k}') \int_{\mathbb{R}^3} S(\mathsf{k}',\mathsf{k})\psi(\mathsf{k})(1-f(t,\mathsf{x},\mathsf{k}))\, d\mathsf{k} \\ &\quad - \int_{\mathbb{R}^3} d\mathsf{k}\, f(t,\mathsf{x},\mathsf{k})\psi(\mathsf{k}) \int_{\mathbb{R}^3} S(\mathsf{k},\mathsf{k}')(1-f(t,\mathsf{x},\mathsf{k}'))\, d\mathsf{k}'\,.\end{aligned}$$

Interchanging k and k' in the first of the above integrals, we infer

$$\begin{aligned}&\int_{\mathbb{R}^3} Q(f)(t,\mathsf{x},\mathsf{k})\,\psi(\mathsf{k})\, d\mathsf{k} \\ &\qquad = \int_{\mathbb{R}^3} d\mathsf{k}\, f(t,\mathsf{x},\mathsf{k}) \int_{\mathbb{R}^3} S(\mathsf{k},\mathsf{k}')(1-f(t,\mathsf{x},\mathsf{k}'))(\psi(\mathsf{k}')-\psi(\mathsf{k}))\, d\mathsf{k}'\,.\end{aligned}$$

Therefore, the proof is achieved by choosing $\psi(\mathsf{k}) = 1$, $\mathsf{k} \in \mathbb{R}^3$. □

4. Space homogeneous solutions: Existence and uniqueness.

In the case when $\mathsf{E} \equiv 0$ and the unknown function f only depends on (t, k), bearing in mind (3.7), the Boltzmann equation (2.1) becomes

$$\frac{\partial f}{\partial t} = (1-f)A(f) - fB(f). \tag{4.1}$$

We look for solutions $f : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}$ to (4.1) that are continuous in $\mathbb{R}_0^+ \times \mathbb{R}^3$, continuously differentiable with respect to the first variable at each point $(t,\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$, and, moreover, fulfill the initial condition

$$f(0,\mathsf{k}) = f_0(\mathsf{k}), \tag{4.2}$$

where $f_0 \in C^0(\mathbb{R}^3)$ and $0 \le f_0(\mathsf{k}) \le 1$ for all $\mathsf{k} \in \mathbb{R}^3$.

Theorem 4.1. *Let assumptions* (a_1) *and* (a_2), *and* (3.6) *be satisfied. Then problem* (4.1)–(4.2) *admits a unique solution f such that* $0 \le f(t,\mathsf{k}) \le 1$ *in* $\mathbb{R}_0^+ \times \mathbb{R}^3$.

Proof. We first define, for every $g \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3)$,

$$H(g)(t,\mathsf{k}) = \exp\left(\int_0^t [A(g)(r,\mathsf{k}) + B(g)(r,\mathsf{k})]\, dr\right), \quad (t,\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3. \tag{4.3}$$

A simple computation shows that the requests on possible solutions f of (4.1)–(4.2) are equivalent to the following: $f \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3)$, $0 \le f(t,\mathsf{k}) \le 1$, and

$$f(t,\mathsf{k}) = \frac{1}{H(f)(t,\mathsf{k})}\left[f_0(\mathsf{k}) + \int_0^t H(f)(s,\mathsf{k})A(f)(s,\mathsf{k})ds\right] \tag{4.4}$$

for $(t, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$.

Choose a real number α complying with

$$\alpha > \frac{1}{2} \max\left\{1, \left[M_A + (1 + M_A)(M_A + M_B)\right]^2\right\}, \tag{4.5}$$

and denote by X the real linear space of all continuous functions $g : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}$ such that

$$\|g\|_\alpha = \sup\left\{\exp\left(-\alpha t^2\right)|g(t, \mathsf{k})| : (t, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3\right\} < +\infty.$$

Arguing as in [6, pp. 2–3], we see that $(X, \|\cdot\|_\alpha)$ is a Banach space. Let

$$Z = \left\{g \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3) : 0 \le g(t, \mathsf{k}) \le 1 \text{ for every } (t, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3\right\}.$$

Clearly, $Z \subseteq X$. For every $g \in Z$, we define

$$T(g)(t, \mathsf{k}) = \frac{1}{H(g)(t, \mathsf{k})}\left[f_0(\mathsf{k}) + \int_0^t H(g)(s, \mathsf{k})A(g)(s, \mathsf{k})ds\right], \quad (t, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3.$$

The function $T(g)$ is continuous in $\mathbb{R}_0^+ \times \mathbb{R}^3$ because, by Proposition 3.1, $A(g)$, $B(g) \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3)$. Further, again owing to Proposition 3.1,

$$\begin{aligned} 0 \le T(g)(t, \mathsf{k}) &= \frac{1}{H(g)(t, \mathsf{k})}\left[f_0(\mathsf{k}) + H(g)(t, \mathsf{k}) - 1 - \int_0^t H(g)(s, \mathsf{k})B(g)(s, \mathsf{k})ds\right] \\ &\le \frac{1}{H(g)(t, \mathsf{k})}\left[f_0(\mathsf{k}) + H(g)(t, \mathsf{k}) - 1\right] \le 1 \end{aligned}$$

for every $(t, \mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$, so $T(g) \in Z$. Hence, $T(Z) \subseteq Z$. The set Z is closed in X and, in view of (4.4), a function $f \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3)$ is a solution to problem (4.1)–(4.2) if and only if it is a fixed point of $T : Z \to Z$. Then, the proof is accomplished by showing that T is a contraction with respect to the norm $\|\cdot\|_\alpha$.

Let g, $h \in Z$. Taking into account (3.1)–(3.3), for every t, $s \in \mathbb{R}_0^+$, $t \ge s$, and every $\mathsf{k} \in \mathbb{R}^3$, one has

$$\begin{aligned} &\left|\exp\left(-\int_s^t \left[A(g)(r, \mathsf{k}) + B(g)(r, \mathsf{k})\right]dr\right) - \exp\left(-\int_s^t \left[A(h)(r, \mathsf{k}) + B(h)(r, \mathsf{k})\right]dr\right)\right| \\ &\quad \le \left|\int_s^t \left[A(g)(r, \mathsf{k}) + B(g)(r, \mathsf{k})\right]dr - \int_s^t \left[A(h)(r, \mathsf{k}) + B(h)(r, \mathsf{k})\right]dr\right| \\ &\quad = \left|\int_s^t \left[A(g - h)(r, \mathsf{k}) - A^*(g - h)(r, \mathsf{k})\right]dr\right| \\ &\quad \le (M_A + M_B)\|g - h\|_\alpha \int_s^t \exp\left(\alpha r^2\right)dr. \end{aligned}$$

Therefore, due to (4.3),

$$\left|\frac{H(g)(s, \mathsf{k})}{H(g)(t, \mathsf{k})} - \frac{H(h)(s, \mathsf{k})}{H(h)(t, \mathsf{k})}\right| \le (M_A + M_B)\|g - h\|_\alpha \int_s^t \exp\left(\alpha r^2\right)dr. \tag{4.6}$$

From this inequality, written for $s = 0$, we obtain

$$\left|\frac{f_0(\mathsf{k})}{H(g)(t, \mathsf{k})} - \frac{f_0(\mathsf{k})}{H(h)(t, \mathsf{k})}\right| \le |f_0(\mathsf{k})|(M_A + M_B)\|g - h\|_\alpha \int_0^t \exp\left(\alpha r^2\right)dr.$$

A simple computation shows that

$$\int_0^t \exp\left(\alpha r^2\right) dr \le \frac{1}{\sqrt{2\alpha}}\, \exp\left(\alpha t^2\right).$$

Consequently,

$$\left|\frac{f_0(\mathsf{k})}{H(g)(t,\mathsf{k})} - \frac{f_0(\mathsf{k})}{H(h)(t,\mathsf{k})}\right| \le \frac{M_A + M_B}{\sqrt{2\alpha}}\, \|g-h\|_\alpha \exp\left(\alpha t^2\right). \tag{4.7}$$

Next, bearing in mind (4.6), for every $(t,\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$ we have

$$\begin{aligned}
&\left|\frac{1}{H(g)(t,\mathsf{k})}\int_0^t H(g)(s,\mathsf{k})A(g)(s,\mathsf{k})ds - \frac{1}{H(h)(t,\mathsf{k})}\int_0^t H(h)(s,\mathsf{k})A(h)(s,\mathsf{k})ds\right| \\
&\quad \le \left|\int_0^t \frac{H(g)(s,\mathsf{k})}{H(g)(t,\mathsf{k})}A(g)(s,\mathsf{k})ds - \int_0^t \frac{H(g)(s,\mathsf{k})}{H(g)(t,\mathsf{k})}A(h)(s,\mathsf{k})ds\right| \\
&\quad + \left|\int_0^t \frac{H(g)(s,\mathsf{k})}{H(g)(t,\mathsf{k})}A(h)(s,\mathsf{k})ds - \int_0^t \frac{H(h)(s,\mathsf{k})}{H(h)(t,\mathsf{k})}A(h)(s,\mathsf{k})ds\right| \\
&\quad \le \int_0^t |A(g)(s,\mathsf{k}) - A(h)(s,\mathsf{k})|\, ds \\
&\quad + (M_A + M_B)\, \|g-h\|_\alpha \int_0^t ds \int_s^t |A(h)(s,\mathsf{k})| \exp\left(\alpha r^2\right) dr \\
&\quad \le M_A\|g-h\|_\alpha \left[\int_0^t \exp\left(\alpha s^2\right) ds + (M_A+M_B)\int_0^t ds \int_s^t \exp\left(\alpha r^2\right) dr\right] \\
&\quad \le M_A\|g-h\|_\alpha \left[\frac{1}{\sqrt{2\alpha}} \exp\left(\alpha t^2\right) + (M_A+M_B)\int_0^t dr \int_0^r \exp\left(\alpha r^2\right) ds\right] \\
&\quad = M_A\|g-h\|_\alpha \left[\frac{1}{\sqrt{2\alpha}} \exp\left(\alpha t^2\right) + (M_A+M_B)\, \frac{1}{2\alpha}\left(\exp\left(\alpha t^2\right) - 1\right)\right] \\
&\quad \le \frac{M_A\left(1 + M_A + M_B\right)}{\sqrt{2\alpha}} \|g-h\|_\alpha \exp\left(\alpha t^2\right)
\end{aligned}$$

because, by (4.5), $2\alpha > 1$. The preceding inequality, together with (4.7), gives

$$\|T(g) - T(h)\|_\alpha \le \frac{M_A + (1+M_A)\left(M_A+M_B\right)}{\sqrt{2\alpha}} \|g-h\|_\alpha.$$

Thus, the desired conclusion immediately follows from (4.5). □

Usually, one requires that the function $\mathsf{k} \to f(t,\mathsf{k})$ be integrable on $\mathbb{R}^3$. In the present setting this is achieved by assuming only that the initial datum f_0 is integrable.

THEOREM 4.2. *Let f_0 and f be as in Theorem* 4.1. *If $f_0 \in L^1(\mathbb{R}^3)$ then the function $\mathsf{k} \to f(t,\mathsf{k})$ is integrable on $\mathbb{R}^3$ for every $t \in \mathbb{R}_0^+$. Furthermore,* $\int_{\mathbb{R}^3} f(t,\mathsf{k})\, d\mathsf{k} = \int_{\mathbb{R}^3} f_0(\mathsf{k})\, d\mathsf{k}$.

Proof. Fix $\tau > 0$. We first note that if $g \in C^0(\mathbb{R}_0^+ \times \mathbb{R}^3) \cap L^1([0,\tau] \times \mathbb{R}^3)$ and $0 \le g(t,\mathsf{k}) \le 1$ for all $(t,\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$, then $A(g)$ is integrable on $[0,\tau] \times \mathbb{R}^3$. In fact, from Proposition 3.2 and (3.9) it follows that

$$\int_0^\tau dt \int_{\mathbb{R}^3} A(g)(t,\mathsf{k})\, d\mathsf{k} \le M_A \int_0^\tau dt \int_{\mathbb{R}^3} g(t,\mathsf{k})\, d\mathsf{k}\ .$$

Since $A(g)$ is nonnegative in $\mathbb{R}_0^+ \times \mathbb{R}^3$, the Tonelli theorem [9, p. 138] implies that $A(g) \in L^1([0,\tau] \times \mathbb{R}^3)$.

We now adopt the notation used in the proof of Theorem 4.1 and define $f_m = T(f_{m-1})$, $m \in \mathbb{N}$. Obviously, the sequence $\{f_m\}$ converges to f in Z. Moreover, we easily have

$$f_m(t,\mathsf{k}) \leq f_0(\mathsf{k}) + \int_0^t A\left(f_{m-1}\right)(s,\mathsf{k})\,ds$$

for every $m \in \mathbb{N}$, $(t,\mathsf{k}) \in \mathbb{R}_0^+ \times \mathbb{R}^3$. Since f_0 is integrable on $[0,\tau] \times \mathbb{R}^3$, the same holds for $A(f_0)$. Arguing by induction, we obtain $f_m \in L^1([0,\tau] \times \mathbb{R}^3)$ for all $m \in \mathbb{N}$ because, in view of (3.9),

$$\int_{\mathbb{R}^3} f_m(t,\mathsf{k})\,d\mathsf{k} \leq \int_{\mathbb{R}^3} f_0(\mathsf{k})\,d\mathsf{k} + M_A \int_0^t ds \int_{\mathbb{R}^3} f_{m-1}(s,\mathsf{k})\,d\mathsf{k}\ , \tag{4.8}$$

and thus

$$\int_0^\tau dt \int_{\mathbb{R}^3} f_m(t,\mathsf{k})\,d\mathsf{k} \leq \tau \left[\int_{\mathbb{R}^3} f_0(\mathsf{k})\,d\mathsf{k} + M_A \int_0^\tau ds \int_{\mathbb{R}^3} f_{m-1}(s,\mathsf{k})\,d\mathsf{k}\right].$$

It is a simple matter to see, by induction again and using inequality (4.8), that

$$\int_{\mathbb{R}^3} f_m(t,\mathsf{k})\,d\mathsf{k} \leq \exp(M_A t) \int_{\mathbb{R}^3} f_0(\mathsf{k})\,d\mathsf{k}$$

for $m \in \mathbb{N}$, $t \in \mathbb{R}_0^+$. Therefore, owing to Fatou's lemma [9, p. 129], the function $\mathsf{k} \to f(t,\mathsf{k})$ is integrable on $\mathbb{R}^3$ for every $t \in \mathbb{R}_0^+$ and f also belongs to $L^1([0,\tau] \times \mathbb{R}^3)$. From (3.7) we infer $Q(f) \in L^1([0,\tau] \times \mathbb{R}^3)$, so $\partial f/\partial t$ is integrable on $[0,\tau] \times \mathbb{R}^3$. Finally, taking into account Proposition 3.2, we get

$$\int_{\mathbb{R}^3} \frac{\partial f(t,\mathsf{k})}{\partial t}\,d\mathsf{k} = \int_{\mathbb{R}^3} Q(f)(t,\mathsf{k})\,d\mathsf{k} = 0 \quad \text{for every } t \in \mathbb{R}_0^+ .$$

Hence,

$$0 = \int_0^\tau dt \int_{\mathbb{R}^3} \frac{\partial f(t,\mathsf{k})}{\partial t}\,d\mathsf{k} = \int_{\mathbb{R}^3} d\mathsf{k} \int_0^\tau \frac{\partial f(t,\mathsf{k})}{\partial t}\,dt = \int_{\mathbb{R}^3} f(\tau,\mathsf{k})\,d\mathsf{k} - \int_{\mathbb{R}^3} f_0(\mathsf{k})\,d\mathsf{k}\ .$$

This completes the proof. □

5. Concluding remarks. We feel that it would be worthwhile to make some observations. The first concerns the presence of the term $\delta(\varepsilon' - \varepsilon \pm \hbar\omega_i)$ inside the collision operator Q. It leads to defining $Q(f)$ and, consequently, to solving problem (4.1)–(4.2) in the setting of functions $(t,\mathsf{k}) \to f(t,\mathsf{k})$ that are at least continuous with respect to the second variable. Actually, since the kernel of Q does not depend on t, the global continuity for f is quite natural.

The second remark concerns the continuity of the function $\mathcal{G}_i$ on the whole space $\mathbb{R}^3 \times \mathbb{R}^3$. In several applications and simulations, functions $\mathcal{G}_i$ defined and continuous only in a proper subset of $\mathbb{R}^3 \times \mathbb{R}^3$ are considered. In any case, $\mathcal{G}_i$ is continuous in the closed set $C = \left\{(\mathsf{k},\mathsf{k}') \in \mathbb{R}^3 \times \mathbb{R}^3 : |\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k})| = \hbar\omega_i\right\}$. By using the Tietze extension theorem [9, p. 192], we are able to find a continuous extension on the whole $\mathbb{R}^3 \times \mathbb{R}^3$. Now, in Appendix C it is shown that the integral

$$\int_{\mathbb{R}^3} \mathcal{G}_i(\mathsf{k},\mathsf{k}') f(t,\mathsf{k}') \delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) \pm \hbar\omega_i)\,d\mathsf{k}'$$

really can be calculated whenever we know the function $\mathcal{G}_i$ only on the set C. Hence, there is no loss of generality in assuming that $\mathcal{G}_i$ is continuous in $\mathbb{R}^3 \times \mathbb{R}^3$.

We next note that, by using hypotheses (a$_1$) and (a$_2$), the collision operator Q is reduced to an integral one, say $\tilde{Q}$, with a continuous kernel (see Appendix C). This new map naturally acts on the space of continuous functions, but it also may be considered in other function spaces. Consequently, the existence of solutions (in the relevant sense) to the full equation

$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_{\mathsf{k}}\varepsilon \cdot \nabla_{\mathsf{x}} f - \frac{e}{\hbar}\mathsf{E}\cdot\nabla_{\mathsf{k}} f = \tilde{Q}(f) \tag{5.1}$$

could be investigated in appropriate settings. Possible solutions of (5.1) do not solve, in general, the unmodified equation (2.1). Nevertheless, like in the case of the classical Boltzmann equation for a perfect rarefied gas, the study of (5.1) is of interest.

The final observation deals with the possibility of studying (2.1) as, in the right-hand side, the term describing electron–electron interactions is added to Q. In the simple parabolic case, it has the following expression [19]:

$$Q_{e-e}(f) = \int_{\mathbb{R}^3} \left[S_{e-e}(f)(\mathsf{k}',\mathsf{k})f'(1-f) - S_{e-e}(f)(\mathsf{k},\mathsf{k}')f(1-f')\right]\, d\mathsf{k}' \ , \tag{5.2}$$

where

$$\begin{aligned} S_{e-e}(f)(\mathsf{k},\mathsf{k}') &= \int_{\mathbb{R}^6} \mathcal{B}(|\mathsf{k}-\mathsf{k}'|, |\mathsf{k}-\mathsf{k}_*|) f_*(1-f'_*)\delta_\varepsilon\, \delta_{\mathsf{k}}\, d\mathsf{k}_*\ d\mathsf{k}'_*\,, \\ \delta_\varepsilon &= \delta\left(\varepsilon(\mathsf{k}) + \varepsilon(\mathsf{k}_*) - \varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}'_*)\right)\ , \\ \delta_{\mathsf{k}} &= \delta(\mathsf{k}+\mathsf{k}_* - \mathsf{k}' - \mathsf{k}'_*)\,, \end{aligned}$$

and $\mathcal{B}$ is a function satisfying suitable conditions. The operator Q_{e-e} seems to have the same form of Q, except for the kernel S_{e-e} that now depends on f. This is not completely true. In fact, $S_{e-e}(f)$ might not be defined even if $\mathcal{B}$ and f are continuous and bounded. This happens, for instance, when $\mathsf{k} = \mathsf{k}'$, in which case $\delta_\varepsilon\, \delta_{\mathsf{k}}$ becomes $\delta(\varepsilon(\mathsf{k}_*) - \varepsilon(\mathsf{k}'_*))\, \delta(\mathsf{k}_* - \mathsf{k}'_*)$.

A similar trouble occurs in the classical Boltzmann equation [4] where, to avoid the singularity, a cutoff is often introduced. The same approach may be tried in the present case. Unfortunately, several further difficulties arise whenever a general expression for ε is maintained.

Appendix A. In this appendix we list the most common expressions of the particle energy ε considered in applications and simulations [3], [10]. It is easily seen that each of them complies with assumptions (a$_1$) and (a$_2$).

$$\varepsilon(\mathsf{k}) = \frac{\hbar^2|\mathsf{k}|^2}{2m} \qquad \text{(parabolic case),}$$

$$\varepsilon(\mathsf{k}) = \varepsilon_c + \frac{\hbar^2}{2m_c}\ |\mathsf{k}-\mathsf{k}_c|^2 \qquad \text{(general parabolic case),}$$

$$\varepsilon(\mathsf{k}) = \frac{\hbar^2}{2}\left(\frac{k_1^2}{m_1} + \frac{k_2^2}{m_2} + \frac{k_3^2}{m_3}\right) \qquad \text{(ellipsoidal model),}$$

$$\varepsilon(\mathsf{k}) = \frac{\hbar^2|\mathsf{k}|^2}{2m}\,(1 - g(\vartheta,\varphi)) \qquad \text{(warped case),}$$

$$\varepsilon(\mathsf{k}) = \frac{-1 + \sqrt{1 + 2\dfrac{\alpha}{m}\hbar^2|\mathsf{k}|^2}}{2\alpha} \qquad \text{(nonparabolic case),}$$

$$\varepsilon(\mathsf{k}) = \frac{-1+\sqrt{1+2\alpha\hbar^2\left(\dfrac{k_l^2}{m_l}+\dfrac{k_t^2}{m_t}\right)}}{2\alpha} \qquad \text{(nonparabolic case)}.$$

Here, m, ε_c, m_c, k_c, m_1, m_2, m_3, α, m_l, and m_t are constant parameters. Moreover, k_1, k_2, k_3 denote the components of the vector k, and k_l, k_t are the lengths of the longitudinal and transverse projections of k with respect to an assigned direction. Finally, the function $g : [0, 2\pi] \times [0, \pi] \to \mathbb{R}$ is continuous and strictly less than one.

Appendix B. A well-known definition of the Dirac distribution δ is given by using the following family of functions with compact support [8].

Let $r \in \mathbb{R}^+$ and let $\omega_r : \mathbb{R} \to \mathbb{R}_0^+$ be the function defined by setting, for $z \in \mathbb{R}$,

$$\omega_r\,(z) = \begin{cases} \dfrac{\exp(1/2)}{K_1(1/2) - K_0(1/2)}\,\dfrac{1}{r}\,\exp\left(-\dfrac{r^2}{r^2-z^2}\right) & \text{if } |z| < r, \\ 0 & \text{otherwise,} \end{cases}$$

where K_0 and K_1 are modified Bessel functions [1], so

$$\int_{-\infty}^{+\infty} \omega_r\,(z)\; dz = 1.$$

The following lemma is an easy extension of useful facts concerning the regularization of a function [2, p. 30].

LEMMA B.1. *Let V be a nonempty subset of a real Euclidean space, let $\lambda \in C^0(\mathbb{R}^3)$, and let $\gamma \in C^0(V \times \mathbb{R}^3 \times \mathbb{R})$. Then, for every $(\mathsf{v}, \mathsf{k}) \in V \times \mathbb{R}^3$, one has*

$$\lim_{r\to 0^+} \int_{-\infty}^{+\infty} \gamma(\mathsf{v}, \mathsf{k}, u)\omega_r\,(u - \lambda(\mathsf{k}))\, du = \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k})). \tag{B.1}$$

Further, (B.1) *uniformly holds on any compact set $\Delta \subseteq V \times \mathbb{R}^3$.*

Proof. The first assertion is well known (see, for instance, [2, Lemma 2.18]). So, let us pick a compact subset Δ of $V \times \mathbb{R}^3$. Since

$$\int_{-\infty}^{+\infty} \gamma(\mathsf{v}, \mathsf{k}, u)\omega_r\,(u - \lambda(\mathsf{k}))\, du = \int_{-\infty}^{+\infty} \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}) + s)\omega_r\,(s)\, ds, \quad (\mathsf{v}, \mathsf{k}) \in \Delta,$$

we get

$$\begin{aligned} &\left|\int_{-\infty}^{+\infty} \gamma(\mathsf{v}, \mathsf{k}, u)\omega_r\,(u - \lambda(\mathsf{k}))\, du - \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}))\right| \\ &\quad\le \int_{-r}^{r} |\gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}) + s) - \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}))|\,\omega_r\,(s)\, ds \\ &\quad\le \sup\left\{|\gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}) + s) - \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}))| : (\mathsf{v}, \mathsf{k}) \in \Delta,\ |s| \le r\right\}. \end{aligned}$$

Then, by using the uniform continuity of the function $(\mathsf{v}, \mathsf{k}, s) \to \gamma(\mathsf{v}, \mathsf{k}, \lambda(\mathsf{k}) + s)$ on the compact set $\Delta \times [-r, r]$, we see that the second assertion is true too. $\square$

Throughout this appendix we denote by W a generic interval of a real Euclidean space and by p and λ two functions belonging to $C^0(W \times \mathbb{R}^3 \times \mathbb{R}^3)$ and $C^0(\mathbb{R}^3)$, respectively. For every $(\mathsf{w}, \mathsf{k}) \in W \times \mathbb{R}^3$, we define

$$\int_{\mathbb{R}^3} p(\mathsf{w}, \mathsf{k}, \mathsf{k}')\, \delta(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))\, d\mathsf{k}' = \lim_{r\to 0^+} \int_{\mathbb{R}^3} p(\mathsf{w}, \mathsf{k}, \mathsf{k}')\omega_r\,(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))\; d\mathsf{k}'. \tag{B.2}$$

Since, in view of (a$_{23}$), the set $\{\mathsf{k}' \in \mathbb{R}^3 : |\varepsilon(\mathsf{k}') - \lambda(\mathsf{k})| < r\}$ is bounded, the function $\mathsf{k}' \to \omega_r(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))$ has compact support. Therefore,

$$\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))\, d\mathsf{k}' < +\infty \quad \text{for all } r > 0.$$

To prove that the right-hand side of (B.2) is a real number, we make a change of variables in the integral. Set $u_0 = \varepsilon(\mathsf{k}_0)$ and $\Omega_0 = [u_0, +\infty[\times D$. Assumption (a$_2$) guarantees that η maps $\mathbb{R}_0^+ \times D$ onto $[u_0, +\infty[$ and, for fixed (ϑ, φ), that the function $\rho \to \eta(\rho, \vartheta, \varphi)$ is strictly increasing on $\mathbb{R}_0^+$. Hence, there exists a unique function $v : \Omega_0 \to \mathbb{R}_0^+$ satisfying

$$\eta(v(u,\vartheta,\varphi),\vartheta,\varphi) = u \quad \text{for every } (u,\vartheta,\varphi) \in \Omega_0. \tag{B.3}$$

Lemma B.2. *The function v is continuous in Ω_0.*

Proof. Pick $(u,\vartheta,\varphi) \in \Omega_0$ and choose a sequence $\{(u_k,\vartheta_k,\varphi_k)\} \subseteq \Omega_0$ converging to (u,ϑ,φ). Moreover, set $\rho_k = v(u_k,\vartheta_k,\varphi_k)$, $k \in \mathbb{N}$. Making use of (B.3) and assumption (a$_{23}$), we see that $\{\rho_k\}$ is bounded. Due to the continuity of η, for any convergent subsequence $\{\rho_{r_k}\}$, one has

$$u = \lim_{k\to\infty} u_{r_k} = \lim_{k\to\infty} \eta(\rho_{r_k}, \vartheta_{r_k}, \varphi_{r_k}) = \eta\left(\lim_{k\to\infty} \rho_{r_k}, \vartheta, \varphi\right).$$

Thus, again by (B.3), $\lim_{k\to\infty} \rho_{r_k} = v(u,\vartheta,\varphi)$. This clearly forces $\lim_{k\to\infty} v(u_k,\vartheta_k,\varphi_k) = v(u,\vartheta,\varphi)$. □

The preceding proof is motivated by the fact that standard implicit function theorems cannot be applied to prove the continuity of the function v on the boundary of the set Ω_0.

Since $v(u_0,\vartheta,\varphi) = 0$, $(\vartheta,\varphi) \in D$, a continuous extension of v over $\Omega = \mathbb{R} \times D$ can be obtained by setting $v(u,\vartheta,\varphi) = 0$ for every $(u,\vartheta,\varphi) \in]-\infty, u_0[\times D$. Finally, if $(u,\vartheta,\varphi) \in \Omega$, we define

$$J(u,\vartheta,\varphi) = \begin{cases} 0 & \text{if } u \le u_0, \\ [v(u,\vartheta,\varphi)]^2 \left[\left.\dfrac{\partial \eta(\rho,\vartheta,\varphi)}{\partial \rho}\right|_{\rho = v(u,\vartheta,\varphi)} \right]^{-1} \sin\varphi & \text{otherwise.} \end{cases}$$

Lemma B.2 and hypothesis (a$_{22}$) imply $J \in C^0(\Omega)$. Moreover, we have the following lemma.

Lemma B.3. *For every $(\mathsf{w},\mathsf{k}) \in W \times \mathbb{R}^3$, one has*

$$\begin{aligned} &\lim_{r\to 0^+} \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))\, d\mathsf{k}' \\ &\qquad = \int_D p(\mathsf{w},\mathsf{k},\mathsf{k}_0 + v(\lambda(\mathsf{k}),\vartheta,\varphi)\mathsf{n}) J(\lambda(\mathsf{k}),\vartheta,\varphi) d\vartheta\, d\varphi. \end{aligned} \tag{B.4}$$

Further, (B.4) *uniformly holds on any compact subset Δ of $W \times \mathbb{R}^3$.*

Proof. Fix $(\mathsf{w},\mathsf{k}) \in W \times \mathbb{R}^3$ and observe that

$$\begin{aligned} &\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r(\varepsilon(\mathsf{k}') - \lambda(\mathsf{k}))\, d\mathsf{k}' \\ &\quad = \int_D d\vartheta\, d\varphi\, \sin\varphi \int_0^{+\infty} p(\mathsf{w},\mathsf{k},\mathsf{k}_0 + \rho\mathsf{n})\omega_r(\varepsilon(\mathsf{k}_0 + \rho\mathsf{n}) - \lambda(\mathsf{k}))\, \rho^2 d\rho, \quad r > 0. \end{aligned}$$

Let us write $\hat{\gamma}(\mathsf{w},\mathsf{k},u,\vartheta,\varphi)$ in place of $p(\mathsf{w},\mathsf{k},\mathsf{k}_0+v(u,\vartheta,\varphi)\mathsf{n})J(u,\vartheta,\varphi)$. By making the change of variable $\rho=v(u,\vartheta,\varphi)$ in the last integral, we get

$$\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}')-\lambda(\mathsf{k})\right)\,d\mathsf{k}' = \int_D d\vartheta\,d\varphi \int_{u_0}^{+\infty} \hat{\gamma}(\mathsf{w},\mathsf{k},u,\vartheta,\varphi)\omega_r\left(u-\lambda(\mathsf{k})\right)du$$

and, due to the properties of the function J,

$$\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}')-\lambda(\mathsf{k})\right)\,d\mathsf{k}' = \int_D d\vartheta\,d\varphi \int_{-\infty}^{+\infty} \hat{\gamma}(\mathsf{w},\mathsf{k},u,\vartheta,\varphi)\,\omega_r\left(u-\lambda(\mathsf{k})\right)du.$$

Moreover, owing to Lemma B.1,

$$\text{(B.5)}\qquad \lim_{r\to 0^+}\int_{-\infty}^{+\infty} \hat{\gamma}(\mathsf{w},\mathsf{k},u,\vartheta,\varphi)\,\omega_r\left(u-\lambda(\mathsf{k})\right)du = \hat{\gamma}(\mathsf{w},\mathsf{k},\lambda(\mathsf{k}),\vartheta,\varphi)$$

uniformly in $(\vartheta,\varphi)\in D$. Hence,

$$\begin{aligned}\lim_{r\to 0^+}&\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}')-\lambda(\mathsf{k})\right)\,d\mathsf{k}'\\ &= \lim_{r\to 0^+}\int_D d\vartheta\,d\varphi \int_{-\infty}^{+\infty} \hat{\gamma}(\mathsf{w},\mathsf{k},u,\vartheta,\varphi)\,\omega_r\left(u-\lambda(\mathsf{k})\right)du\\ &= \int_D \hat{\gamma}(\mathsf{w},\mathsf{k},\lambda(\mathsf{k}),\vartheta,\varphi)d\vartheta\,d\varphi.\end{aligned}$$

Since, by Lemma B.1 again, (B.5) uniformly holds on $\Delta\times D$ for any compact subset Δ of $W\times\mathbb{R}^3$, the conclusion follows. □

The preceding lemma, together with (B.2), leads to

$$\text{(B.6)}\qquad \begin{aligned}&\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\lambda(\mathsf{k}))\,d\mathsf{k}'\\ &\quad= \int_D p(\mathsf{w},\mathsf{k},\mathsf{k}_0+v(\lambda(\mathsf{k}),\vartheta,\varphi)\mathsf{n})J(\lambda(\mathsf{k}),\vartheta,\varphi)\,d\vartheta\,d\varphi.\end{aligned}$$

Consequently, the continuity theorem for integrals depending on a parameter yields the following lemma.

Lemma B.4. *The function*

$$(\mathsf{w},\mathsf{k})\to\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta\left(\varepsilon(\mathsf{k}')-\lambda(\mathsf{k})\right)\,d\mathsf{k}',\quad (\mathsf{w},\mathsf{k})\in W\times\mathbb{R}^3,$$

is continuous in $W\times\mathbb{R}^3$.

We denote by $C_0^0(\mathbb{R}^3)$ the space of all functions $\psi\in C^0(\mathbb{R}^3)$ which have compact support suppψ.

Lemma B.5. *Let* $\mu\in\mathbb{R}$. *Then, for every* $\psi\in C_0^0(\mathbb{R}^3)$, $\mathsf{w}\in W$,

$$\text{(B.7)}\qquad \begin{aligned}&\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\\ &\quad= \int_{\mathbb{R}^3} d\mathsf{k}'\int_{\mathbb{R}^3} \psi(\mathsf{k})p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\,.\end{aligned}$$

Proof. Choose $\psi \in C_0^0(\mathbb{R}^3)$ and $\mathsf{w} \in W$. By Lemma B.4, the left-hand side of (B.7) is a real number. Since, owing to Lemma B.3,

$$\lim_{r\to 0^+} \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu\right)\, d\mathsf{k}' = \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}'$$

uniformly in $\{\mathsf{w}\} \times \mathrm{supp}\psi$, we get

$$\begin{aligned}\text{(B.8)} \qquad & \int_{\mathbb{R}^3} d\mathsf{k}\, \psi(\mathsf{k}) \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}' \\ &\quad = \lim_{r\to 0^+} \int_{\mathbb{R}^3} d\mathsf{k} \int_{\mathbb{R}^3} \psi(\mathsf{k}) p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu\right)\, d\mathsf{k}'\,.\end{aligned}$$

For every $r > 0$ the function $(\mathsf{k},\mathsf{k}') \to \psi(\mathsf{k})p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu\right)$ has compact support in $\mathbb{R}^3 \times \mathbb{R}^3$ because, by assumption (a_{23}), the set

$$\left\{(\mathsf{k},\mathsf{k}') \in \mathbb{R}^3 \times \mathbb{R}^3 : \mathsf{k} \in \mathrm{supp}\psi,\ |\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu| \le r\right\}$$

is compact. Therefore, due to (B.8) and Lemma B.3 again,

$$\begin{aligned}& \int_{\mathbb{R}^3} d\mathsf{k}\, \psi(\mathsf{k}) \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}' \\ &\quad = \lim_{r\to 0^+} \int_{\mathbb{R}^3} d\mathsf{k}' \int_{\mathbb{R}^3} \psi(\mathsf{k}) p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu\right)\, d\mathsf{k} \\ &\quad = \int_{\mathbb{R}^3} d\mathsf{k}' \lim_{r\to 0^+} \int_{\mathbb{R}^3} \psi(\mathsf{k}) p(\mathsf{w},\mathsf{k},\mathsf{k}')\omega_r\left(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu\right)\, d\mathsf{k} \\ &\quad = \int_{\mathbb{R}^3} d\mathsf{k}' \int_{\mathbb{R}^3} \psi(\mathsf{k}) p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}\,.\end{aligned}$$

This completes the proof. □

We note that the presence of the Dirac distribution δ prevents the establishment of (B.7) through the usual formulas on the change of order of integration. As an example, if $p(z,z') = (1 + |z||z'-1|)/(1+z^2)$, $z, z' \in \mathbb{R}$, then

$$\int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} p(z,z')\,\delta(z'-1)\, dz' = \int_{-\infty}^{+\infty} p(z,1)\, dz = \pi\,,$$

whereas the other iterated integral does not exist, because

$$\int_{-\infty}^{+\infty} p(z,z')\, dz = +\infty \quad \text{whenever } z' \neq 1\,.$$

Lemma B.6. *Let $\mu \in \mathbb{R}$. If $p(\mathsf{w},\mathsf{k},\mathsf{k}') \ge 0$ in $W \times \mathbb{R}^3 \times \mathbb{R}^3$ and, for any $\mathsf{w} \in W$, the function*

$$\mathsf{k} \to \int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k}) - \mu)\, d\mathsf{k}'$$

is integrable on $\mathbb{R}^3$, then (B.7) *holds for all nonnegative bounded $\psi \in C^0(\mathbb{R}^3)$.*

Proof. Pick $w \in W$ and choose a function ψ with the asserted properties. Moreover, let $\psi_n : \mathbb{R} \to \mathbb{R}_0^+$, $n \in \mathbb{N}$, be defined as in (3.8). Assumption (a_{23}) guarantees that the function $\mathsf{k} \to \psi_n(\varepsilon(\mathsf{k}))$ has compact support. Hence, by Lemma B.5,

$$\begin{aligned}\text{(B.9)}\qquad &\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\psi_n(\varepsilon(\mathsf{k}))\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\\ &\quad=\int_{\mathbb{R}^3} d\mathsf{k}'\int_{\mathbb{R}^3}\psi(\mathsf{k})\psi_n(\varepsilon(\mathsf{k}))p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\end{aligned}$$

for $n \in \mathbb{N}$. Both sides of (B.9) are nonnegative and less than or equal to the real number

$$\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\,.$$

Since $\psi_n(z) \le \psi_{n+1}(z)$ for all $n \in \mathbb{N}$ and $z \in \mathbb{R}$, the sequence of functions

$$\mathsf{k} \to \psi(\mathsf{k})\psi_n(\varepsilon(\mathsf{k}))\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'$$

is nondecreasing and pointwise convergent in $\mathbb{R}^3$ to

$$\mathsf{k} \to \psi(\mathsf{k})\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\,.$$

So, the monotone convergence theorem [9, p. 129] gives

$$\begin{aligned}\text{(B.10)}\qquad &\lim_{n\to\infty}\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\psi_n(\varepsilon(\mathsf{k}))\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\\ &\quad=\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\,.\end{aligned}$$

Making use of (B.6), it is not difficult to see that

$$\begin{aligned}&\int_{\mathbb{R}^3} d\mathsf{k}'\int_{\mathbb{R}^3}\psi(\mathsf{k})\psi_n(\varepsilon(\mathsf{k}))p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\\ &\quad=\int_{\mathbb{R}^3} d\mathsf{k}'\,\psi_n(\varepsilon(\mathsf{k}')-\mu)\int_{\mathbb{R}^3}\psi(\mathsf{k})p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\end{aligned}$$

for every $n \in \mathbb{N}$. This formula, together with (B.9) and (B.10), leads to

$$\begin{aligned}&\int_{\mathbb{R}^3} d\mathsf{k}\,\psi(\mathsf{k})\int_{\mathbb{R}^3} p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}'\\ &\quad=\lim_{n\to\infty}\int_{\mathbb{R}^3} d\mathsf{k}'\,\psi_n(\varepsilon(\mathsf{k}')-\mu)\int_{\mathbb{R}^3}\psi(\mathsf{k})p(\mathsf{w},\mathsf{k},\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})-\mu)\,d\mathsf{k}\,.\end{aligned}$$

Now, the same arguments used to prove (B.10) yield the conclusion. □

Appendix C. In view of (B.6), we get

$$\begin{aligned}\text{(C.1)}\qquad &\int_{\mathbb{R}^3}\mathcal{G}_i(\mathsf{k},\mathsf{k}')f(t,\mathsf{k}')\delta(\varepsilon(\mathsf{k}')-\varepsilon(\mathsf{k})\pm\hbar\omega_i)\,d\mathsf{k}'\\ &\quad=\int_D \mathcal{G}_i(\mathsf{k},\xi_\mp(\vartheta,\varphi,\mathsf{k}))f(t,\xi_\mp(\vartheta,\varphi,\mathsf{k}))J(\varepsilon(\mathsf{k})\mp\hbar\omega_i,\vartheta,\varphi)d\vartheta\,d\varphi,\end{aligned}$$

where $\xi_{\mp}(\vartheta,\varphi,\mathsf{k}) = \mathsf{k}_0 + v(\varepsilon(\mathsf{k}) \mp \hbar\omega_i, \vartheta, \varphi)\mathsf{n}$. Now $\mathsf{k}' = \xi_{\mp}(\vartheta,\varphi,\mathsf{k})$ satisfies the equation $\varepsilon(\mathsf{k}') = \varepsilon(\mathsf{k}) \mp \hbar\omega_i$ for all $(\vartheta,\varphi) \in D$, $\mathsf{k} \in \mathbb{R}^3$. In fact, recalling the definition of the function η (see hypothesis (a_2)) and (B.3), we have

$$\begin{aligned}\varepsilon(\mathsf{k}') &= \varepsilon\left(\mathsf{k}_0 + v(\varepsilon(\mathsf{k}) \mp \hbar\omega_i, \vartheta, \varphi)\mathsf{n}\right)\\ &= \eta\left(v(\varepsilon(\mathsf{k}) \mp \hbar\omega_i, \vartheta, \varphi), \vartheta, \varphi\right) = \varepsilon(\mathsf{k}) \mp \hbar\omega_i.\end{aligned}$$

Therefore, in the integral at the left-hand side of (C.1), the function $\mathcal{G}_i$ is actually evaluated for (k,k') in the set $\left\{(\mathsf{k},\mathsf{k}') \in \mathbb{R}^3 \times \mathbb{R}^3 : |\varepsilon(\mathsf{k}') - \varepsilon(\mathsf{k})| = \hbar\omega_i\right\}$.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1972.

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

[3] J. S. Blakemore, *Semiconductor Statistics*, Dover, New York, 1987.

[4] C. Cercignani, *The Boltzmann Equation and its Applications*, Springer-Verlag, Berlin, New York, 1988.

[5] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases*, 3rd ed., Cambridge University Press, Cambridge, 1970.

[6] C. Corduneanu, *Integral Equations and Stability of Feedback Systems*, Academic Press, New York, 1973.

[7] R. J. Di Perna and P. L. Lions, *On the Cauchy problem for the Boltzmann equations: Global existence and weak stability*, Ann. of Math., 130 (1989), pp. 321–366.

[8] I. M. Gel'fand and G. E. Shilov, *Generalized Functions*, Vol. I, Academic Press, New York, 1964.

[9] N. B. Haaser and J. A. Sullivan, *Real Analysis*, Dover, New York, 1991.

[10] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, Berlin, New York, 1989.

[11] A. Majorana, *Space homogeneous solutions of the Boltzmann equation describing electron-phonon interactions in semiconductors*, Transport Theory Statist. Phys., 20 (1991), pp. 261–279.

[12] A. Majorana, *Conservation laws from the Boltzmann equation describing electron-phonon interaction in semiconductors*, Transport Theory Statist. Phys., 22 (1993), pp. 81–93.

[13] A. Majorana, *Equilibrium solutions of the non-linear Boltzmann equation for an electron gas in a semiconductor*, Nuovo Cimento B, 108 (1993), pp. 871–877.

[14] F. J. Mustieles, *Global existence of solutions for the nonlinear Boltzmann equation of semiconductor physics*, Rev. Mat. Iberoamericana, 6 (1990), pp. 43–59.

[15] F. J. Mustieles, *Global existence of weak solutions for a system of nonlinear Boltzmann equations of semiconductor physics*, Math. Methods Appl. Sci., 14 (1991), pp. 139–153.

[16] P. A. Markowich, F. Poupaud, and C. Schmeiser, *Diffusion approximation of nonlinear electron phonon collision mechanisms*, RAIRO Model. Math. Anal. Numer., 29 (1995), pp. 857–869.

[17] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Berlin, New York, 1990.

[18] A. Nouri and F. Poupaud, *Stationary solutions of boundary value problems for Maxwell Boltzmann system modeling degenerate semiconductors*, SIAM J. Math. Anal., 26 (1995), pp. 1143–1156.

[19] F. Poupaud, *On a system of nonlinear Boltzmann equations of semiconductor physics*, SIAM J. Appl. Math., 50 (1990), pp. 1593–1606.

[20] H. Smith and H. Højgaard Jensen, *Transport Phenomena*, Clarendon Press, Oxford, 1989.

[21] Th. Vogelsang and W. Hänsch, *A novel approach for including band-structure effects in a Monte Carlo simulation of electron transport in silicon*, J. Appl. Phys., 70 (1991), pp. 1493–1499.

SIAM J. Math. Anal.
Vol. 28, No. 6, pp. 1309–1316, November 1997

ESTIMATES ON THE HEAT KERNEL OF PARABOLIC EQUATIONS WITH ADVECTION*

ADRIAN T. HILL†

Abstract. The paper considers heat kernels of second-order parabolic equations in $\mathbb{R}^N$, with constant uniform diffusion and advective coefficients bounded in the maximum norm. Two critical cases, corresponding to upper and lower solutions, are identified, and explicit solutions are constructed for them in terms of the error function. They are shown to bound above and below all other heat kernels satisfying the same constraints on their advective coefficients by using a method of proof which relates two heat kernels together in a way which resembles the classical parametrix construction. Sharp bounds on the corresponding parabolic solution operators in $L_1(\mathbb{R}^N)$ are obtained as a consequence.

Key words. heat kernels, non-self-adjoint operators

AMS subject classifications. 35B45, 35K05

PII. S003614109630104X

1. Introduction.

Upper and lower pointwise bounds on the heat kernels of second-order parabolic operators were first derived, for the general case of variable diffusion and advection coefficients in L_p spaces, by Aronson [1]. These initial qualitative estimates have more recently been sharpened, in the self-adjoint case, by Davies [2] and Fabes and Stroock [3]. For the non-self-adjoint case of equations with advection, sharper estimates have been obtained by Norris and Stroock [6].

Here, we consider the heat kernel $\Gamma(x,\,t;\,y,\,s)$ of the equation

$$Lu(x,t) \equiv \frac{\partial u}{\partial t} - \Delta\ u + a(x,\,t).\nabla u = 0, \qquad (x,\,t) \in \mathbb{R}^N \times (0,\,\infty), \tag{1.1}$$

with constant uniform diffusion, and aim to obtain sharper estimates in this physically significant special case. Our methods are classically based and, thus, we initially take a to obey assumption (A):

(A) $a(x,\,t)$ is a continuous function, Hölder continuous in x (exponent α), uniformly with respect to $(x,\,t)$ in subsets of the form $\mathbb{R}^N \times [0,\,T)$, for $T > 0$. The a_i also satisfy (1.2).

$$\text{ess. sup}_{(x,\,t)\in\mathbb{R}^N\times[0,\,\infty)}|a_i(x,\,t)| \le M_i \qquad \text{for } M_i \ge 0,\ i = 1,\,\ldots,\,N. \tag{1.2}$$

We recall that the classical heat kernel $\Gamma(x,\,t;\,y,\,s)$, $x,\,y \in \mathbb{R}^N$, $t > s \ge 0$, may be characterized as a solution of (1.1), twice continuously differentiable in space and once in time, such that

$$\lim_{t\searrow s}\int_{\mathbb{R}^N}\Gamma(x,\,t;\,y,\,s)f(y)\,dy = f(x) \tag{1.3}$$

for all continuous $f \in L_1(\mathbb{R}^N)$.

*Received by the editors March 25, 1996; accepted for publication July 30, 1996.
http://www.siam.org/journals/sima/28-6/30104.html

†School of Mathematical Sciences, University of Bath, BA2 7AY, United Kingdom (ath@maths.bath.ac.uk).

To derive a pointwise upper bound on $\Gamma(x,\, t;\, y, s)$ for fixed y and s, we explicitly construct a function $G_M(x-y,\, t-s)$, which is the solution of

$$\frac{\partial u}{\partial t} - \Delta u + \sum_{i=1}^{N} M_i \mathrm{sign}[x_i - y_i] \frac{\partial u}{\partial x_i} = 0, \qquad (x,\, t) \in \mathbb{R}^N \times (0,\, \infty), \tag{1.4}$$

such that

$$\lim_{t \searrow s} \int_{\mathbb{R}^N} G_M(x-y,\, t-s) f(y)\, dy = f(x) \tag{1.5}$$

for all continuous $f \in L_1(\mathbb{R}^N)$. Here, we use the convention $\mathrm{sign}[0] = 0$, and $M = [M_1,\, M_2,\, \ldots,\, M_N]^T$, where the M_i are as in (1.2).

The constructed function $G_M(x-y,\, t-s)$ also satisfies

$$\mathrm{sign}[x_i - y_i] \frac{\partial G_M}{\partial x_i} \leq 0, \qquad i = 1,\, \ldots,\, N, \tag{1.6}$$

and is, therefore, one of the solutions of the nonlinear equation

$$\frac{\partial u}{\partial t} - \Delta u - \sum_{i=1}^{N} M_i \left| \frac{\partial u}{\partial x_i} \right| = 0, \qquad (x,\, t) \in \mathbb{R}^N \times (0,\, \infty). \tag{1.7}$$

We remark that the maximality of nonsingular solutions of (1.7) has previously been considered by Pucci [7].

If one ignores the singularity at $(x,\, t) = (y,\, s)$, then intuitively $G_M(x-y,\, t-s)$ should bound $\Gamma(x,\, t;\, y, s)$ pointwise above, because it is an upper solution for (1.1), (1.3) in the sense of the comparison principle. In our main result we prove this relationship by representing Γ as the sum of G_M and a nonpositive integral term; our representation resembles, in a way, the classical parametrix construction of the heat kernel. A lower bound is similarly obtained by considering $G_{-M}(x,\, t;\, y,\, s)$, the solution of (1.4), (1.5) with $-M_i$ replacing M_i.

By appealing to the machinery of Aronson [1], this classical result is extended to the case where a obeys assumption (B):

$$a \in L_\infty[\mathbb{R}^N \times [0,\, \infty)] \text{ and } a \text{ satisfies (1.2).} \tag{B}$$

The pointwise upper limit on $\Gamma(x,\, t;\, y,\, s)$ is actually attained within the wider class of equations obeying assumption (B). This is seen from the fact that $G_M(x - y, t-s)$ satisfies (1.4) and (1.5) and is, therefore, equal to $\Theta_M(z;\, x,\, t;\, y,\, s)$, the heat kernel of

$$\frac{\partial u}{\partial t} - \Delta u + \sum_{i=1}^{N} M_i \mathrm{sign}[x_i - z_i] \frac{\partial u}{\partial x_i} = 0 \tag{1.8}$$

when $y = z$. A similar argument shows that the lower bound is also attained. Since Aronson [1] has shown that Θ_M may be pointwise approximated by the heat kernels of equations satisfying assumption (A), the upper and lower bounds given by $G_{\pm M}(x,\, t;\, y,\, s)$ are sharp (though not attained) in that case.

In section 4, we use the pointwise upper and lower bounds on $\Gamma(x,\, t;\, y,\, s)$ to deduce sharp upper and lower L_1 bounds on the solutions of (1.1) with $u(0) = u_0 \in L_1(\mathbb{R})$. In particular, we obtain

$$\|u(t)\|_1 \le \prod_{i=1}^{N} \left(2 + M_i^2 t\right) \|u_0\|_1.$$

2. Main result. For $M \in \mathbb{R}$, consider the function $g_M : \mathbb{R} \times (0,\, \infty) \to (0,\, \infty)$ given by

$$g_M(x,\, t) = \frac{1}{2\sqrt{\pi t}} \exp\left\{-\frac{(|x| - Mt)^2}{4t}\right\} + \frac{M}{4} \operatorname{erfc}\left(\frac{|x|}{2\sqrt{t}} - \frac{M\sqrt{t}}{2}\right), \tag{2.1}$$

where $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt$. By direct differentiation,

$$\begin{aligned}
\frac{\partial g_M}{\partial x} &= \frac{-x}{4\sqrt{\pi t^3}} \exp\left\{-\frac{(|x| - Mt)^2}{4t}\right\},\\
\frac{\partial^2 g_M}{\partial x^2} &= \frac{1}{4\sqrt{\pi t^3}} \left(-1 + \frac{x^2}{2t} - \frac{M|x|}{2}\right) \exp\left\{-\frac{(|x| - Mt)^2}{4t}\right\},\\
\frac{\partial g_M}{\partial t} &= \frac{1}{4\sqrt{\pi t^3}} \left(-1 + \frac{x^2}{2t} + \frac{M|x|}{2}\right) \exp\left\{-\frac{(|x| - Mt)^2}{4t}\right\}.
\end{aligned} \tag{2.2}$$

Each of the right-hand sides is Lipschitz continuous. (Note that $\partial^3 g_M/\partial x^3$ is discontinuous at $x = 0$.) We deduce that g_M is a classical solution of

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + M \operatorname{sign}[x] \frac{\partial u}{\partial x} = 0, \qquad (x,\, t) \in \mathbb{R} \times (0,\, \infty). \tag{2.3}$$

We define

$$G_M(x,\, t) = \prod_{i=1}^{N} g_{M_i}(x_i,\, t) \qquad \text{for } M \in \mathbb{R}^N. \tag{2.4}$$

By (2.2)–(2.4), $G_M(x - y,\, t)$ satisfies (1.4) and (1.6).

THEOREM 2.1. *Suppose that for some $M \in [0,\, \infty)^N$, $a(x,\, t)$ satisfies assumption* (B). *Then, for all $x,\, y \in \mathbb{R}^N$, $t > s \ge 0$,*

$$G_{-M}(x - y,\, t - s) \le \Gamma(x,\, t;\, y, s) \le G_M(x - y,\, t - s). \tag{2.5}$$

For each $y \in \mathbb{R}^N$, these bounds are attained for all $x \in \mathbb{R}^N$, $t > s \ge 0$ by the heat kernels $\Theta_{\pm M}(z;\, x,\, t;\, y,\, s)$ of (1.8) *when $y = z$.*

Remark. Aronson [1, Theorem 10, p. 679] shows that $\Gamma(x,\, t;\, y,\, s) = \tilde{\Gamma}(y,\, s;\, x,\, t)$ for $t > s \ge 0$, where $\tilde{\Gamma}$ is the heat kernel for the adjoint problem to (1.1):

$$\tilde{L}[u] \equiv \frac{\partial v}{\partial s} + \Delta v + \nabla.(a(y,\, s)v) = 0, \qquad (y,\, s) \in \mathbb{R}^N \times [0,\, t). \tag{2.6}$$

Hence, Theorem 2.1 also implies pointwise bounds for $\tilde{\Gamma}$.

3. Proof of main result.

We begin here with the following lemma.

LEMMA 3.1. *If* (1.1) *obeys assumption* (A) *for some* $M \in [0, \infty)^N$*, then*

$$\Gamma(x, t; y, s) = G(x-y, t-s) - \int_s^t \int_{\mathbb{R}^N} \Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma, \tag{3.1}$$

where $G(x, t)$ *is either* $G_M(x, t)$ *or* $G_{-M}(x, t)$.

Proof. The proof relies on the account given by Friedman [4, Chapter 1] of the parametrix method. Here, for $t_0 \in (s, t)$ and $D = \{z \mid |z - x| \le 1\}$, the integral on the right-hand side of (3.1) is split up as $J_1 + J_2 + J_3$, where

$$J_1 = \int_s^{t_0} \int_{\mathbb{R}^N} \Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma,$$

$$J_2 = \int_{t_0}^{t} \int_{\mathbb{R}^N \setminus D} \Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma,$$

$$J_3 = \int_{t_0}^{t} \int_{D} \Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma.$$

For J_1 and J_2, $\Gamma(x, t; \eta, \sigma)$ has second derivatives in x, continuous with respect to $(x, t; \eta, \sigma)$. On the other hand, for C and c, generic positive constants,

$$\begin{aligned} |LG(\eta - y, \sigma - s)| &= \left| \sum_{i=1}^N \left(a_i(\eta, \sigma) \mp M_i \text{sign}[\eta_i - y_i] \frac{\partial G}{\partial \eta_i} \right) \right| \\ &\le C \sum_{i=1}^N \left| \frac{\partial G}{\partial \eta_i} \right| \\ &\le \frac{C|\eta - y|}{(\sigma - s)^{(N+2)/2}} \exp\left(-(|\eta - y| - M(\sigma - s))^2 / 4(\sigma - s) \right) \\ &\le \frac{C}{(\sigma - s)^{3/4} |\eta - y|^{(N-1/2)}} \exp\left(-c|\eta - y|^2/(\sigma - s) \right) \end{aligned}$$

and so is absolutely integrable. Hence, differentiation in x and t commutes with integration in J_1 and J_2. (We note here that differentiation of the upper limit t in J_2 is valid, but that this yields a term which one may show to be 0.)

For J_3, we observe that for fixed y and s,

$$LG(\eta - y, \sigma - s) = \sum_{i=1}^N (a_i(\eta, \sigma) \mp M_i \text{sign}[\eta_i - y_i]) \frac{\partial G}{\partial \eta_i}(\eta - y, \sigma - s)$$

is Hölder continuous in η, uniformly for $(\eta, \sigma) \in \mathbb{R}^N \times [t_0, t]$, since (considering (2.2) and (2.4))

$$\frac{\partial G}{\partial \eta_i}(\eta - y, \sigma - s) \qquad \text{and} \qquad \text{sign}[\eta_i - y_i] \frac{\partial G}{\partial \eta_i}(\eta - y, \sigma - s)$$

are both uniformly Lipschitz continuous and assumption (A) implies that $a(\eta, \sigma)$ is uniformly Hölder continuous.

According to Friedman [4, Theorems 1.3–1.5], the first and second derivatives in x commute with integration in J_3, and

$$\frac{\partial J_3}{\partial t} = \int_{t_0}^{t} \int_D \frac{\partial \Gamma(x, t; \eta, \sigma)}{\partial t} LG(\eta - y, \sigma - s)\, d\eta d\sigma + LG(x - y, t - s).$$

The continuity of these derivatives follows from considerations similar to [4, Theorem 1.8, p. 19].

Let $J(x, t; y, s) = J_1 + J_2 + J_3$. From (2.2), (2.4) it follows that $G(x-y, t-s)$ is twice continuously differentiable in x and once in t for $x, y \in \mathbb{R}^N$ and $t > s \geq 0$. Thus, the same property holds for $G(x-y, t-s) + J(x, t; y, s)$.

Combining the above results, we deduce that

$$\begin{aligned} &LG(x-y, t-s) - L\int_s^t \int_{\mathbb{R}^N} \Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma \\ &\quad = LG(x-y, t-s) - LG(x-y, t-s) \\ &\quad - \int_s^t \int_{\mathbb{R}^N} L\Gamma(x, t; \eta, \sigma) LG(\eta - y, \sigma - s)\, d\eta d\sigma = 0. \end{aligned}$$

We must now prove that

$$\lim_{t \searrow s} \int_{\mathbb{R}^N} (G(x-y, t-s) + J(x, t; y, s)) f(y)\, dy = f(x) \tag{3.2}$$

for continuous $f \in L_1(\mathbb{R}^N)$. The contribution from J may be shown to be 0 by following the proof in [4, p. 20]. On the other hand, for fixed $T > 0$, it is readily shown that

$$\begin{aligned} &\left| G(x-y, t-s) - \frac{\exp(-|x-y|^2/(4(t-s)))}{(4\pi(t-s))^{N/2}} \right| \\ &\quad \leq \frac{C}{(t-s)^{(N-1)/2}} \exp\left(\frac{-c|x-y|^2}{(t-s)} \right) \\ &\quad \leq \frac{C}{(t-s)^{1/2}|x-y|^{(N-2)}} \exp\left(\frac{-c|x-y|^2}{(t-s)} \right) \end{aligned}$$

for all $(t-s) \in (0, T]$. Hence,

$$\lim_{t \searrow s} \int_{\mathbb{R}^N} \left(G(x-y, t-s) - \frac{\exp(-|x-y|^2/(4(t-s)))}{(4\pi(t-s))^{N/2}} \right) f(y)\, dy = 0,$$

so (3.2) follows from the well-known properties of the Gaussian kernel.

We deduce that $G(x-y, t-s) + J(x, t; y, s)$ satisfies the characterizing equations (1.1) and (1.3) and is, therefore, equal to $\Gamma(x, t; y, s)$. □

Proof of Theorem 2.1. We first consider (3.1) under assumption (A). It is well known that $\Gamma(x, t; \xi, \sigma) > 0$. On the other hand, for $M \in [0, \infty)^N$,

$$LG_M(x-y, t-s) = \sum_{i=1}^N (a_i(x, t) - M_i \text{sign}[x_i - y_i]) \frac{\partial G_M}{\partial x_i} \geq 0$$

by (1.2) and (1.6). So, (3.1) implies that $\Gamma(x, t; y, s) \leq G_M(x-y, t-s)$. The lower bound follows similarly.

Under conditions which include (1.1) under assumption (B), Aronson [1] defines the weak fundamental solution $\Gamma(x, t; y, s)$, which coincides with the classical heat kernel when coefficients are sufficiently smooth. He considers [1, p. 681] a sequence of classical heat kernels $\Gamma^k(x, t; y, s)$, $k \in \mathbb{N}$, for the problem

$$u_t - \Delta u + a^k(x, t) \cdot \nabla u = 0 \tag{3.3}$$

for $|x| < k$, $t > 0$, with Dirichlet boundary conditions

$$u(x,\, t) = 0, \qquad |x| = k,\ t > 0,$$

and shows that $\Gamma^k(x,\, t;\, y,\, s) \to \Gamma(x,\, t;\, y,\, s)$ pointwise, as $k \to \infty$. Here, $a^k(x,\, t)$ is an integral average of $a(x,\, t)$ formed with a smooth kernel whose support lies in $|x|^2 + t^2 < k^{-2}$. Therefore, $a^k(x,\, t)$ satisfies assumption (A).

Let $H^k(x,\, t;\, y,\, s)$ be the heat kernel of (3.3) in the extended domain $\mathbb{R}^N \times (0,\, \infty)$. Then $\Gamma^k(x,\, t;\, y,\, s) = H^k(x,\, t;\, y,\, s) + v^k(x,\, t;\, y,\, s)$ for $|x|$, $|y| \le k$, where $v^k(x,\, t;\, y,\, s)$ is the interior solution of (3.3) such that

$$v^k(x,\, t;\, y,\, s) = -H^k(x,\, t;\, y,\, s), \qquad |x| = k,\ t > 0.$$

Since H^k satisfies the upper and lower bounds (2.5), it is clear from the parabolic maximum principle [4, Chapter 2] that $v^k(x,\, t;\, y,\, s) \to 0$ pointwise, as $k \to \infty$, for fixed $(x,\, t;\, y,\, s)$. Hence, $\Gamma(x,\, t;\, y,\, s)$ satisfies (2.5).

The last part of the theorem is completed by following the argument given in the introduction. □

4. L^1 bounds. Aronson [1, section 9] shows, under conditions which include the case of (1.1) with $a(x,\, t)$ satisfying assumption (B), that if $u_0 \in L_1(\mathbb{R}^N)$, then

$$u(x,\, t) = \int_{\mathbb{R}^N} \Gamma(x,\, t;\, y,\, 0) u_0(y)\, dy \tag{4.1}$$

is the unique weak solution of the equation in $L_2[[\delta,\, T];\, L_2(\mathbb{R}^N)]$ for any $T > \delta > 0$ and that $u(x,\, t)$ also satisfies the continuity condition

$$\lim_{t \to 0} \|u(t) - u_0\|_1 = 0.$$

Here, we use the representation formula (4.1) and the pointwise bounds (2.5) on $\Gamma(x,\, t;\, y,\, 0)$ to obtain an upper bound on $\|u(t)\|_1 / \|u_0\|_1$.

Since u_0, and hence $u(t)$, may be split as the sum of nonnegative and nonpositive parts, it is sufficient to consider nonnegative u_0 for the purposes of obtaining L_1 bounds. In this case, Fubini's theorem implies that

$$\begin{aligned} \|u(t)\|_1 &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \Gamma(x,\, t;\, y,\, 0) u_0(y)\, dy\, dx \\ &= \int_{\mathbb{R}^N} \left(\int_{\mathbb{R}^N} \Gamma(x,\, t;\, y,\, 0)\, dx \right) u_0(y)\, dy, \qquad t > 0. \end{aligned}$$

From (2.5) we deduce that, provided u_0 is not null,

$$\inf_{y \in \mathbb{R}^N} \int_{\mathbb{R}^N} G_{-M}(x - y,\, t)\, dx \le \frac{\|u(t)\|_1}{\|u_0\|_1} \le \sup_{y \in \mathbb{R}^N} \int_{\mathbb{R}^N} G_M(x - y,\, t)\, dx.$$

That there is in fact no y dependence involved is apparent on a change of variable in the integrals, and one obtains

$$\int_{\mathbb{R}^N} G_{-M}(x,\, t)\, dx \le \frac{\|u(t)\|_1}{\|u_0\|_1} \le \int_{\mathbb{R}^N} G_M(x,\, t)\, dx. \tag{4.2}$$

Remark. Suppose that we take a sequence of smooth nonnegative initial data $u_0^k(x)$, $k \in \mathbb{N}$, for (1.8) such that $\|u_0^k\|_1 = 1$ and the support of u_0^k lies in the set

$|x-z|<k^{-1}$. Let the corresponding solution be denoted $u^k(x, t)$. Fixing $t>0$ and noting that the heat kernel $\Theta_M(z;\, x,\, t;\, y,\, 0)$ of (1.8) is equal to $G_M(x-z,\, t)$ when $y=z$, we deduce that

$$\lim_{k\to\infty}\frac{\|u^k(t)\|_1}{\|u_0^k\|_1}=\int_{\mathbb{R}^N} G_M(x,\, t)\, dx.$$

If one exchanges M_i for $-M_i$, a similar argument shows that the lower bound in (4.2) is also sharp.

The construction (2.4) of G_M implies that

$$\int_{\mathbb{R}^N} G_{\pm M}(x,\, t)\, dx=\prod_{i=1}^{N}\int_{\mathbb{R}} g_{\pm M_i}(x_i,\, t)\, dx_i. \tag{4.3}$$

A calculation using the identity [5, Formula 7.2.5, p. 299]

$$\int_x^\infty \operatorname{erfc}(t)\, dt=-x\operatorname{erfc}(x)+\frac{1}{\sqrt{\pi}}\exp(-x^2)$$

yields

$$\int_{\mathbb{R}} g_M(x,\, t)\, dx=h\left(\frac{M\sqrt{t}}{2}\right), \qquad M\in\mathbb{R},\ t\geq 0, \tag{4.4}$$

where we define

$$h(x)=\left(1+2x^2\right)\operatorname{erfc}(-x)+\frac{2x}{\sqrt{\pi}}\exp\left(-x^2\right). \tag{4.5}$$

The above analysis is summarized in the following theorem.

THEOREM 4.1. *Suppose that L obeys assumption* (B) *and that non-null initial data $u_0\in L_1(\mathbb{R})$ are given for* (1.1). *Then, if u_0 is nonnegative, the solution $u(t)$ satisfies*

$$\prod_{i=1}^{N} h\left(\frac{-M_i\sqrt{t}}{2}\right)\leq\frac{\|u(t)\|_1}{\|u_0\|_1}\leq\prod_{i=1}^{N} h\left(\frac{M_i\sqrt{t}}{2}\right). \tag{4.6}$$

For general $u_0\in L_1(\mathbb{R}^N)\setminus\{0\}$, $u(t)$ obeys the upper bound of (4.6).

For small x, the behavior of $h(x)$ is as seen in Gautschi [5, Formula 7.1.6, p. 297], where for $x\in\mathbb{R}$,

$$\begin{aligned} h(x)&=(1+2x^2)+\frac{2\exp(-x^2)}{\sqrt{\pi}}\left(x+(1+2x^2)\sum_{n=0}^{\infty}\frac{2^n}{1\cdot 3\cdots(2n+1)}x^{2n+1}\right)\\ &=1+\frac{4x}{\sqrt{\pi}}+2x^2+\frac{4x^3}{3\sqrt{\pi}}+\cdots. \end{aligned}$$

When $M_i\sqrt{t}\ll 1$ for each i, the upper bound in (4.6) grows like

$$1+\frac{2\sum_{i=1}^{N} M_i}{\sqrt{\pi}}t^{1/2}, \tag{4.7}$$

while the decay of the lower bound is found by changing the sign of the M_i's.

Using an asymptotic expansion for erfc(x) in negative powers of $|x|$ [5, Formula 7.1.23, p. 298], for $x < 0$ one obtains

$$\begin{aligned} h(x) &= -\frac{2|x|}{\sqrt{\pi}}\exp(-x^2) + (1+2x^2)\mathrm{erfc}(|x|) \\ &\sim \frac{\exp(-x^2)}{\sqrt{\pi}|x|}\left(-2x^2 + (1+2x^2)\left(1 + \sum_{n=1}^{\infty}(-1)^n \frac{1\cdot 3\cdots(2n-1)}{(2x^2)^n}\right)\right). \end{aligned}$$

(If the series is truncated, the error, $h(x)-$R.H.S., is equal to $\theta(x)$ times the first neglected term, where $\theta(x) \in (0, 1)$.) Hence, when $M_i\sqrt{t} \gg 1$ for each i, the lower bound in (4.6) decays like

$$\left(\frac{8}{\pi^{1/2}\overline{M}^3}\right)^N t^{-3N/2}\exp\left(-\sum_{i=1}^{N}\frac{M_i^2 t}{4}\right) \qquad \text{for} \quad \overline{M} \equiv \left(\prod_{i=1}^{N} M_i\right)^{1/N}. \tag{4.8}$$

Since erfc$(x)+$ erfc$(-x) = 2$, (4.5) implies that $h(x) = 2 + 4x^2 - h(-x)$. Thus, when $M_i\sqrt{t} \gg 1$ for each i, the upper bound in (4.6) may be approximated by

$$\prod_{i=1}^{N}(2 + M_i^2 t) \tag{4.9}$$

with very little error. This last result may be attributed to the fact that

$$G_M(x, t) \approx \prod_{i=1}^{N}\frac{M_i}{2} \qquad \text{for } -M_i t < x_i < M_i t,\ i = 1, 2, \ldots, N, \tag{4.10}$$

and rapidly vanishes outside this box.

Remark. Considering non-null nonnegative endpoint data $u(t) = u_0 \in L_\infty(\mathbb{R}^N)$ for the adjoint problem (2.6), a duality argument leads to the bounds

$$\prod_{i=1}^{N} h\left(\frac{-M_i\sqrt{t-s}}{2}\right) \le \frac{\|u(s)\|_\infty}{\|u_0\|_\infty} \le \prod_{i=1}^{N} h\left(\frac{M_i\sqrt{t-s}}{2}\right), \qquad s \in [0, t]. \tag{4.11}$$

Acknowledgments. The author wishes to thank Professor J. F. Toland for suggesting this problem and Dr. J. R. Norris, Dr. S. C. Harris, and Professor L. C. G. Rogers for materially helpful discussions.

REFERENCES

[1] D. G. Aronson, *Non-negative solutions of linear parabolic equations*, Ann. Sci. Norm. Sup. Pisa, 22 (1968), pp. 607–694.

[2] E. B. Davies, *Heat kernels and spectral theory*, Cambridge Tracts in Mathematics 92, Cambridge University Press, Cambridge, UK, 1989.

[3] E. Fabes and D. W. Stroock, *A new proof of Moser's Harnack inequality using the old ideas of Nash*, Arch. Rational Mech. Anal., 96 (1986), pp. 327–338.

[4] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[5] W. Gautschi, *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun, eds., National Bureau of Standards, Washington D.C., 1964.

[6] J. R. Norris and D. W. Stroock, *Estimates on the fundamental solution to heat flows with uniformly elliptic coefficients*, Proc. London Math. Soc. (3), 62 (1991), pp. 373–402.

[7] C. Pucci, *Operatori ellitici estremanti*, Ann. Mat. Pura Appl. (4), 72 (1966), pp. 141–170.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1317–1353, November 1997

004

SPATIAL PATTERNS DESCRIBED BY THE EXTENDED FISHER–KOLMOGOROV EQUATION: PERIODIC SOLUTIONS[*]

L. A. PELETIER[†] AND W. C. TROY[‡]

Abstract. Stationary antisymmetric single-bump periodic solutions of a fourth-order generalization of the Fisher–Kolmogorov (FK) equation are analyzed. The coefficient $\gamma > 0$ of the additional fourth-order spatial derivative is found to be a critical parameter. If $\gamma \leq \frac{1}{8}$, the family of periodic solutions is still very similar to that of the FK equation. However, if $\gamma > \frac{1}{8}$, it is possible to distinguish different families of periodic solutions and the structure of such solutions is much richer.

Key words. differential equations, nonlinear, periodic solutions, phase transitions

AMS subject classifications. 34C15, 34C25, 35Q35

PII. S0036141095280955

1. Introduction. In this paper we shall study the formation of spatially periodic patterns in bistable systems described by the *extended Fisher–Kolmogorov* (EFK) equation

$$\frac{\partial u}{\partial t} = -\gamma \frac{\partial^4 u}{\partial x^4} + \frac{\partial^2 u}{\partial x^2} + u - u^3, \qquad \gamma > 0. \tag{1.1}$$

This equation was proposed in 1987 by Coullet, Elphick, and Repaux [8] and in 1988 by Dee and van Saarloos [11] as a generalization of the classical Fisher–Kolmogorov (FK) equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u - u^3, \tag{1.2}$$

which has played an important role in the studies of pattern formation in bistable systems (cf. [3, 13, 14, 16, 20, 25, 26]). The term "bistable" refers here to the fact that the uniform states $u = \pm 1$ are stable as solutions of the equation

$$\frac{du}{dt} = u - u^3.$$

The EFK equation arises in the study of singular points (so-called Lifshitz points [14]) in phase transitions and as the evolution equation in gradient systems described by the energy functional

$$I(u) = \int \left\{ \frac{\gamma}{2}(u'')^2 + \frac{\beta}{2}(u')^2 + F(u) \right\} dx, \qquad \gamma > 0,\ \beta \in \mathbf{R}, \tag{1.3}$$

where F denotes the double-well potential

$$F(u) = \frac{1}{4}(1 - u^2)^2. \tag{1.4}$$

[*] Received by the editors February 3, 1995; accepted for publication (in revised form) August 15, 1996. This research was supported by NSF grant DMS9002028.

http://www.siam.org/journals/sima/28-6/28095.html

[†] Mathematical Institute, Leiden University, Leiden, The Netherlands (peletier@wi.leidenuniv.nl).

[‡] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (troy@vms.cis.pitt.edu).

We then obtain the EFK equation when we choose $\beta = 1$. Another important application of this equation is found in the theory of instabilities in nematic liquid crystals [5, 27].

In studies of second-order materials [10, 17, 18] one also finds the functional $I(u)$. Here $\beta < 0$. The stationary points of $I(u)$ are then equivalent to the equilibrium solutions of the Swift–Hohenberg equation

$$\frac{\partial u}{\partial t} = -\Big(1 + \frac{\partial^2}{\partial x^2}\Big)^2 u + \alpha\, u - u^3, \qquad \alpha > 0 \tag{1.5}$$

when $\alpha > 1$ through a simple scaling of x, t, and u (see, for instance, [7, 9] and the references therein).

In this paper we are interested in stationary spatial patterns which can be described by the EFK equation and in particular in *periodic patterns.* Thus, we are concerned with bounded solutions $u(x)$ of the equation

$$\gamma u^{iv} = u'' + u - u^3 \quad \text{on } \mathbf{R}. \tag{1.6}$$

Our main objectives are to determine the effect of the added higher-order gradient term and the value of the coefficient γ, on the class of possible stationary periodic patterns, and the qualitative properties of these patterns.

For convenience we recall below the types of stationary patterns that can be described by the FK equation. However, before doing so we introduce some notation. We observe that equation (1.6) has a constant of integration: if u is a solution of equation (1.6), then

$$\mathcal{E}(u) \overset{\text{def}}{=} 2\gamma u' u''' - \gamma (u'')^2 - (u')^2 + \frac{1}{2}(1-u^2)^2 \;=\; \text{constant} \;\overset{\text{def}}{=}\; \frac{\mu}{2}. \tag{1.7}$$

To eliminate arbitrary shifts, we always place the origin at a zero of u:

$$u(0) = 0 \tag{1.8}$$

whenever a zero exists.

For the special value $\gamma = 0$, equation (1.6) reduces to the stationary FK equation

$$u'' + u - u^3 = 0. \tag{1.9}$$

PROPOSITION 1.1. *The FK equation* ($\gamma = 0$) *has the following bounded stationary solutions.*

(a) *There exists a unique solution u of equation* (1.9) *(a kink), such that*

$$(u, u') \to (\pm 1, 0) \quad \text{as } x \to \pm\infty.$$

It is antisymmetric and monotone and is given explicitly by

$$u(x) = \tanh\Big(\frac{x}{\sqrt{2}}\Big).$$

This solution corresponds to the values $\gamma = 0$ and $\mu = 0$ in (1.7).

(b) *For each $\mu \in (0, 1)$, there exists a unique periodic solution u of* (1.9) *such that $u'(0) > 0$. It is* (i) *antisymmetric with respect to its zeros,* (ii) *symmetric with respect* *to the location of its maxima and its minima,* (iii) *concave where it is positive and convex where it is negative, and*

$$\text{(iv)} \qquad \max\{|u(x)| : x \in \mathbf{R}\} = \sqrt{1 - \sqrt{\mu}}.$$

(c) *There exist no bounded periodic solutions of* (1.9) *when* $\mu \notin (0, 1)$ *and no bounded solutions of* (1.9) *when* $\mu \notin [0, 1]$.

In two earlier papers [21, 24] we investigated the existence and properties of kinks of the EFK equation. They are solutions $u(x)$ of equation (1.6) which have the properties

$$(u, u', u'', u''') \to (\pm 1, 0, 0, 0) \quad \text{as } x \to \pm\infty. \tag{1.10}$$

We found that kinks exist for every $\gamma > 0$ but that their character and number change abruptly at $\gamma = \frac{1}{8}$. When $\gamma \le \frac{1}{8}$, there exists a unique odd monotone kink with the same characteristic properties as the kink of the FK equation. However, we have recently shown that for $\gamma > \frac{1}{8}$, there exists a countably infinite number of odd kinks, and none of these is monotone in its approach to ±1 as $x \to \pm\infty$ [22]. We collect these results in the following proposition.

Proposition 1.2. (a) *For each* $\gamma > 0$*, there exists an odd solution of equation* (1.6) *which satisfies* (1.10).

(b) *If* $\gamma \le \frac{1}{8}$*, there exists one and only one kink which is odd and monotone.*

(c) *If* $\gamma > \frac{1}{8}$*, there exists for every integer* $n \ge 0$ *a kink with* $2n + 1$ *zeros.*

The critical value $\gamma = \frac{1}{8}$ is related to the linearization of (1.6) around $u = 1$ or $u = -1$. For $\gamma \le \frac{1}{8}$ the corresponding eigenvalues are all real, whilst for $\gamma > \frac{1}{8}$ they are all complex.

In the present paper we investigate stationary *periodic* solutions of the EFK equation and we shall inquire how parts (b) and (c) of Proposition 1.1 generalize when we take $\gamma > 0$. We shall restrict this study to periodic solutions u which (*i*) are *odd*, (ii) are *symmetric* with respect to the location of their extrema, and (iii) have a single relative maximum between consecutive zeros. Thus, let ζ be the *first* positive zero of u'. Then we are concerned with periodic solutions of (1.6) with the following properties:

$$\begin{aligned} u(-x) &= -u(x) && \text{for } x \in \mathbf{R}, \\ u(\zeta - y) &= u(\zeta + y) && \text{for } y \in \mathbf{R}. \end{aligned} \tag{1.11}$$

We refer to solutions which satisfy (1.11) as *single-bump* periodic solutions. In this paper we shall only discuss such periodic solutions.

As we do with the FK equation we only consider solutions for which

$$\mathcal{E}(u) \ge 0 \quad \text{or} \quad \mu \ge 0.$$

We conjecture that if $\gamma > \frac{1}{8}$, there exist periodic solutions for some *negative* values of μ as well. However, we leave their analyses to a further study.

In our first result, we find that when $0 < \mu < 1$ the periodic solutions of the FK equation continue to exist for all $\gamma > 0$.

Theorem A. *Let* $0 < \mu < 1$ *and* $\gamma > 0$*. Then there exists a periodic solution* $u(x)$ *of* (1.6) *such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} < 1.$$

When $\mu \geq 1$ the FK equation has no periodic solutions and we see below that this also continues to be true for all $\gamma > 0$.

Theorem B. *Let $\mu \geq 1$ and $\gamma > 0$. Then there exists no periodic solution of equation* (1.6).

When $\mu = 0$ and $\gamma > \frac{1}{8}$ the situation becomes very different from that of the FK equation. Whereas the FK equation has no periodic solutions for this value of μ, the EFK equation has two branches of periodic solutions bifurcating from the unique odd monotone kink $U(x)$ at $\gamma = \frac{1}{8}$. This is the content of the following two theorems.

Theorem C. *Let $\mu = 0$.*

(a) *If $0 < \gamma \leq \frac{1}{8}$, then there exist no periodic solutions.*

(b) *If $\gamma > \frac{1}{8}$, then there exist a periodic solution $u_1(x)$ such that*

$$\max\{|u_1(x)| : x \in \mathbf{R}\} < 1$$

and a periodic solution $u_2(x)$ such that

$$\max\{|u_2(x)| : x \in \mathbf{R}\} \in (1, \sqrt{2}).$$

Theorem D. *Let $\mu = 0$ and let $\{\gamma_i\}$ be a sequence such that*

$$\gamma_i \searrow \frac{1}{8} \quad \textit{as } i \to \infty.$$

For each $i \geq 1$, let u_i be a periodic solution corresponding to γ_i. Then

$$u_i(x) \to U(x) \quad \textit{as } i \to \infty,$$

where U is the unique odd monotone kink corresponding to $\gamma = \frac{1}{8}$. The convergence is uniform on compact intervals.

We conjecture that in addition to the single-bump periodic solutions of Theorem C, infinitely many multibump periodic solutions bifurcate from U at $\gamma = \frac{1}{8}$.

A result which is analogous to Theorem D holds for periodic solutions when $0 < \mu < 1$ and $\gamma \to 0$ (cf. Lemma 6.3). They converge to the periodic solution of the FK equation for the given value of μ.

Periodic solutions with amplitude larger than 1 continue to exist when $\gamma > \frac{1}{8}$ and $\mu > 0$ is sufficiently small. However, if either $\gamma \leq \frac{1}{8}$ or $\mu \geq \frac{4}{9}$, they cannot exist.

Theorem E. *Let $\mu \geq 0$ and $\gamma > 0$. There exist no periodic solutions $u(x)$ such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} > 1$$

when one of the following conditions is satisfied:

(a) $0 < \gamma \leq \frac{1}{8}$,

(b) $\mu \geq \frac{4}{9}$.

For the global behavior of the branches of periodic solutions we obtain several bounds. We begin with a universal upper bound.

Theorem F. *Let $0 \leq \mu < 1$ and $\gamma > 0$. Then any periodic solution satisfies*

$$|u(x, \gamma)| < \sqrt{2} \quad \textit{for } x \in \mathbf{R}.$$

We also prove the following lower bound.

THEOREM G. *Let $0 \le \mu < 1$ and $\gamma > (\frac{2}{5})^4$. Then any periodic solution satisfies*

$$\max\{|u(x,\gamma)| : x \in \mathbf{R}\} > \frac{1}{50}\sqrt{\frac{1-\mu}{\log 2}}.$$

Let u be a periodic solution which satisfies (1.11). Then for its slope $u'(0)$ at the origin we prove an upper and a lower bound:

$$\frac{1}{5}\sqrt{1-\mu} < \gamma^{1/4}u'(0,\gamma) < \{8(1-\mu)\log 2\}^{1/4} \tag{1.12}$$

for $\gamma > \frac{1}{8}$ if $\mu = 0$ and for $\gamma > (\frac{2}{5})^4$ if $0 < \mu < 1$. These bounds enable us to obtain information about the behavior of periodic solutions for large values of γ. The description of their limiting behavior involves the reduced problem

$$\begin{cases} v^{iv} = v - v^3, & \text{(1.13a)} \\ v(0) = 0, \quad v''(0) = 0, & \text{(1.13b)} \\ v'(0) = \omega, \quad v'''(0) = -\dfrac{1-\mu}{4\omega} & \text{(1.13c)} \end{cases}$$

in which ω is a positive number.

THEOREM H. *Let $0 \le \mu < 1$. Suppose that $\{\gamma_i\}$ is a sequence which tends to infinity and $\{u_i\}$ is a sequence of periodic solutions which satisfy* (1.11). *Then there exist a subsequence, which we also denote by $\{\gamma_i\}$, and a periodic solution V of problem* (1.13) *such that*

$$u_i(\gamma_i^{1/4}s,\gamma_i) \to V(s) \quad \textit{as } i \to \infty$$

uniformly on compact sets.

Outline of the shooting method. We now give a brief description of the topological shooting method used to prove the existence of the families of odd periodic solutions described in Theorem A and Theorem C(b).

Since we are looking for odd solutions of (1.6), it is sufficient to consider the equation on $\mathbf{R}^+$ only and with initial conditions of the form

$$\big(u(0), u'(0), u''(0), u'''(0)\big) = (0, \alpha, 0, \beta), \tag{1.14}$$

where α and β are real. It is easily verified that a solution u of (1.6) satisfying (1.14) must be odd, and that $-u$ is also a solution. We shall see that α cannot be zero, and so we may choose α to be positive. It follows from (1.7) that

$$\beta = \frac{1}{2\alpha\gamma}\Big\{\alpha^2 - \frac{1-\mu}{2}\Big\}. \tag{1.15}$$

In Theorem A we assume that γ and μ are fixed, with $\gamma > 0$ and $0 < \mu < 1$. We are then free to vary α, our shooting parameter. We begin our analysis by showing that for each $\alpha > 0$ there exists a finite value $\xi(\alpha) > 0$, which depends continuously on α, such that

$$u'(x,\alpha) > 0 \quad \text{for } 0 < x < \xi(\alpha) \qquad \text{and} \qquad u'(\xi(\alpha),\alpha) = 0. \tag{1.16}$$

We shall show that $u'''(\xi(\alpha),\alpha) < 0$ for small values of $\alpha > 0$, that there is an $\tilde{\alpha} > 0$ for which $u'''(\xi(\tilde{\alpha}),\tilde{\alpha}) > 0$, and that

$$0 < u(\xi(\alpha),\alpha) < \sqrt{1-\sqrt{\mu}} \quad \text{for } 0 < \alpha < \tilde{\alpha}. \tag{1.17}$$

From these observations and the continuity of $\xi(\alpha)$ with respect to α we conclude that there is an intermediate value $\alpha_- \in (0, \tilde{\alpha})$ such that $u'''(\xi(\alpha_-), \alpha_-) = 0$. At $\alpha = \alpha_-$ we have

$$0 < u(\xi(\alpha_-), \alpha_-) < \sqrt{1 - \sqrt{\mu}}, \quad u'(\xi(\alpha_-), \alpha_-) = 0, \quad \text{and} \quad u'''(\xi(\alpha_-), \alpha_-) = 0,$$

and it follows from (1.7) and the definition of $\xi(\alpha_-)$ that $u''(\xi(\alpha_-), \alpha_-) < 0$. By reflecting the graph $\{(x, u(x, \alpha_-)) : 0 \leq x \leq \xi(\alpha_-)\}$ with respect to the endpoints $x = 0$ and $x = \xi(\alpha_-)$, we obtain the desired periodic solution. The details of the analysis described above are given in sections 2 and 3.

In section 5 we turn to the proof of Theorem C(b) in which we set $\mu = 0$ and restrict γ to satisfy $\gamma > \frac{1}{8}$. For $\gamma \in (0, \frac{1}{8}]$, part (a) of Theorem C states that no periodic solutions exist, and this result is proved in section 4. However, as γ passes through $\frac{1}{8}$ from below, a linearization of (1.6) around the constant solution $u = 1$ shows that the eigenvalues change from real to complex, two of the eigenvalues have positive real part and two have negative real part. Because of this change in character of the eigenvalues, Coullet, Elphick, and Repaux [8] conjectured that the range $\gamma > \frac{1}{8}$ is where one would expect to observe complex pattern formation. Such patterns could include families of periodic and aperiodic solutions, multibump heteroclinic and homoclinic orbits (kinks, respectively, solitons), and possibly chaos.

One of the main goals of our investigation of the EFK equation is to resolve this conjecture of Coullet, Elphick, and Repaux and to determine the different types of solutions that exist for the parameter regime $\mu = 0$ and $\gamma > \frac{1}{8}$. In Theorem C(b) we state our first existence result for this range of parameters. We prove that there are at least two families of periodic solutions; one of these is characterized by the fact that the relative maxima all lie *below* $u = 1$, while the relative maxima of the second family all lie *above* $u = 1$. In addition we show that both families of periodic solutions bifurcate from the odd, monotone kink (the heteroclinic orbit connecting $u = -1$ and $u = +1$) at $\gamma = \frac{1}{8}$ and continue to exist for all $\gamma > \frac{1}{8}$.

The proof of Theorem C(b) is based on the same shooting technique we used for the proof of Theorem A. Again, for each $\alpha > 0$ we find that there exists a first $\xi(\alpha) \in \mathbf{R}^+$ for which $u'(\xi(\alpha), \alpha) = 0$. In the proof of Theorem A we found that $u(\xi(\alpha), \alpha) \neq 1$ because $\mu \neq 0$. However, now we have $\mu = 0$ and indeed we find that there is a critical value $\tilde{\alpha}$ for which $u(\xi(\tilde{\alpha}), \tilde{\alpha}) = 1$. It follows from (1.6), (1.7), and uniqueness that

$$(1.18) \quad u(\xi(\tilde{\alpha}), \tilde{\alpha}) = 1, \quad u'(\xi(\tilde{\alpha}), \tilde{\alpha}) = 0, \quad u''(\xi(\tilde{\alpha}), \tilde{\alpha}) = 0, \quad \text{and} \quad u'''(\xi(\tilde{\alpha}), \tilde{\alpha}) > 0.$$

To proceed with our shooting argument we need to know that $\xi(\alpha)$ is continuous. When $u'' \neq 0$ this is an easy consequence of the implicit function theorem, but when $u'' = 0$, as it is at $\xi(\tilde{\alpha})$, this is no longer obvious and the situation is much more delicate. In Lemma 5.8 we further refine a method originally developed in [21], which uses u as an independent variable, to prove this important property.

Having established continuity of $\xi(\alpha)$ we conclude that there exists an $\alpha_- \in (0, \tilde{\alpha})$ for which $u'''(\xi(\alpha_-), \alpha_-) = 0$. As before, it follows that $u(\cdot, \alpha_-)$ is a periodic solution and by construction its relative maxima lie *below* $u = 1$.

For our second periodic solution, whose relative maxima are all greater than 1, we show that for $\alpha > \tilde{\alpha}$ sufficiently large,

$$u(\xi(\alpha), \alpha) > 1 \quad \text{and} \quad u'''(\xi(\alpha), \alpha) < 0.$$

Remembering that $u'''(\xi(\tilde{\alpha}),\tilde{\alpha}) > 0$, we conclude from the continuity of $\xi(\alpha)$ that there exists an $\alpha_+ > \tilde{\alpha}$ for which

$$u(\xi(\alpha_+),\alpha_+) > 1, \quad u'(\xi(\alpha_+),\alpha_+) = 0, \quad \text{and} \quad u'''(\xi(\alpha_+),\alpha_+) = 0.$$

Again, we conclude that $u(\cdot,\alpha_+)$ is a periodic solution, but now the construction ensures that its relative maxima lie *above* $u = 1$.

We emphasize that the key result in the proof of the existence of both families of periodic orbits is Lemma 5.8, which implies that $\xi(\alpha)$ is continuous for all $\alpha > 0$ if $\mu = 0$ and $\gamma > \frac{1}{8}$. We have recently extended this property [23] and proved that all subsequent relative maxima $\xi_i(\alpha)$ and minima $\eta_i(\alpha)$, $i = 1, 2, \ldots$, have the same continuity property. In turn this has allowed us to further refine our shooting method to establish the existence of complicated types of solutions, such as multibump heteroclinic orbits [22] and chaotic patterns [23]. The topological shooting argument developed here and extended in [22] and [23] enables one to prove chaos without having to verify the typical transversality condition (see, for instance, [12]) required by the dynamical systems approach. This verification can be very difficult.

Preliminary results suggest that the method developed here presents a framework which can be used to investigate stationary spatial patterns described by a large class of model equations, such as the Swift–Hohenberg equation (1.5), and equations describing soliton solutions in nonlinear optical fibers [1], traveling waves in a suspension bridge [19], and the deflections of an asymmetrically supported strutt:

$$u^{iv} + Pu'' + u - u^2 = 0, \qquad P \in \mathbf{R}. \tag{1.19}$$

In a series of papers [2, 6, 4] this equation has been studied by completely different (Hamiltonian) methods. Although the nonlinearity in this equation is not cubic, and the emphasis here lies on homoclinic orbits, it is interesting to compare the two different methods and the different types of results they generate. Finally, we note that in [15] a variational approach in combination with a partition of function spaces into topological subclasses has proved to be successful in analyzing equations such as (1.6) and (1.9).

2. Preliminaries. Our basic method for proving existence of odd periodic solutions is a shooting technique, so we consider the initial value problem

$$\begin{cases} \gamma u^{iv} = u'' + u - u^3, \qquad x > 0, & \text{(2.1a)} \\ u(0) = 0,\ u'(0) = \alpha,\ u''(0) = 0,\ u'''(0) = \beta. & \text{(2.1b)} \end{cases}$$

We may restrict our attention to $\alpha > 0$; if $\alpha = 0$, then (1.7) implies that $\mu = 1$ and we shall show in Lemma 2.2 that in this case no periodic solution which satisfies (1.11) can exist. Thus, if we fix $\mu \geq 0$ and $\gamma > 0$, then (1.7) yields

$$\beta = \beta(\alpha) = \frac{1}{2\gamma\alpha}\Big\{\alpha^2 - \frac{1-\mu}{2}\Big\}. \tag{2.2}$$

We seek a positive value of α such that the solution $u(x,\alpha)$ has the properties (1.11). That is,

$$u'(x,\alpha) > 0 \quad \text{for } 0 \leq x < \xi, \tag{2.3a}$$

$$u'(\xi,\alpha) = 0 \quad \text{and} \quad u'''(\xi,\alpha) = 0 \tag{2.3b}$$

for some finite $\xi = \xi(\alpha) > 0$. It is easily verified that a solution defined on $[0,\xi]$, which satisfies (2.1)–(2.3), can be extended to yield a periodic solution of period 4ξ. Thus, we define

$$\xi(\alpha) = \sup\{x > 0 : u'(\cdot,\alpha) > 0 \ \text{ on } \ [0,x)\}.$$

In this section we will show that for all values $\alpha > 0$, except those for which the corresponding solution u is a monotone kink, $\xi(\alpha)$ is finite and that u is bounded on $[0,\xi]$ with horizontal slope at ξ. Therefore, to satisfy (2.3) we must still determine $\alpha > 0$ so that

$$u'''(\xi(\alpha),\alpha) = 0.$$

At times we shall find it convenient to adopt a different formulation for the initial value problem (2.1). Since we construct periodic solutions from strictly monotone segments defined on $[0,\xi]$, we may introduce u as an independent variable, as was done in [21] for the study of kinks. Denoting the inverse function of $u(x)$ by $x(u)$, we set

$$t = u \quad \text{and} \quad z(t) = (u')^2(x(t)). \tag{2.4}$$

This yields

$$z'(t) = 2u''(x) \quad \text{and} \quad z''(t) = 2\frac{u'''(x)}{u'(x)}. \tag{2.5}$$

Hence, upon substitution into (1.7), we obtain

$$\begin{cases} zz'' = \dfrac{(z')^2}{4} + \dfrac{1}{\gamma}\{z - f_\mu(t)\}, \quad t > 0, & \text{(2.6a)} \\ z(0) = \alpha^2 \quad \text{and} \quad z'(0) = 0, & \text{(2.6b)} \end{cases}$$

where

$$f_\mu(t) = \frac{1}{2}\{(t^2-1)^2 - \mu\}. \tag{2.7}$$

We denote the solution by $z(t,\alpha)$ and write

$$\tau(\alpha) = \sup\{t > 0 : z(\cdot,\alpha) > 0 \ \text{ on } \ [0,t)\}.$$

From (2.4) and the definition of $\xi(\alpha)$ it follows that

$$\tau(\alpha) = \lim_{x\to\xi(\alpha)^-} u(x,\alpha).$$

To be assured that $z(\cdot,\alpha)$ corresponds to a periodic solution $u(\cdot,\alpha)$, we need to prove the existence of a positive α for which

$$0 < \xi(\alpha) < \infty, \quad 0 < \tau(\alpha) < \infty, \tag{2.8a}$$

and

$$\lim_{t\to\tau(\alpha)^-} z(t,\alpha) = 0, \quad \lim_{t\to\tau(\alpha)^-} \sqrt{z(t,\alpha)}z''(t,\alpha) = 0. \tag{2.8b}$$

Then (2.8), together with (2.4) and (2.5), implies that $u(\cdot,\alpha)$ satisfies (2.3) so that u is periodic.

LEMMA 2.1. *Suppose that $\mu \geq 0$ and $\gamma > 0$.*

(a) *For any $\alpha \in \mathbf{R}^+$, we have*

$$u(\xi(\alpha),\alpha) < \infty \quad \textit{and} \quad u'(\xi(\alpha),\alpha) = 0. \tag{2.9}$$

(b) *If $\mu > 0$, then*

$$\xi(\alpha) < \infty \quad \textit{for any } \alpha \in \mathbf{R}^+.$$

(c) *If $\mu = 0$, then*

$$\xi(\alpha) < \infty \quad \begin{cases} \textit{for any } \alpha \in \mathbf{R}^+ & \textit{if } \gamma > \dfrac{1}{8}, \\ \textit{for any } \alpha \in \mathbf{R}^+ \setminus \{\alpha_0\} & \textit{if } 0 < \gamma \leq \dfrac{1}{8}. \end{cases}$$

Here $\alpha_0 = U'(0)$, and U is the unique odd monotone kink found in [21].

Proof. (a) We write (2.6a) as

$$(z^{3/4})'' = \frac{3}{4\gamma}\frac{z-f}{z^{5/4}}, \tag{2.10}$$

where we have suppressed the subscript μ from f, and we define

$$\tau_0 = \sup\{t \in (0,\tau) : z' < 0 \ \text{ on } \ (0,t)\},$$

if $z' < 0$ in a right-neighborhood of the origin, and $\tau_0 = 0$ otherwise.

We distinguish two cases:

$$\text{(i)} \ \ \tau_0 = \tau \quad \text{and} \quad \text{(ii)} \ \ \tau_0 < \tau.$$

(i) In this case, $z(t) < \alpha^2$ for $0 < t < \tau$. Suppose that $\tau = \infty$ (i.e., $u(\xi(\alpha),\alpha) = \infty$). Then, since $f(t) \sim \frac{1}{2}t^4$ as $t \to \infty$, there exists a $T > 0$ such that (2.10) reduces to

$$(z^{3/4})'' < -1 \quad \text{for } t > T,$$

which implies that $\tau < \infty$, a contradiction. Thus, it must be the case that $u(\xi(\alpha),\alpha) < \infty$. If $\lim_{t\to\tau^-} z(t) > 0$, then standard theory applied to (2.6) shows that z continues to exist, with $z > 0$ on an interval $[\tau, \tau+\varepsilon)$, contradicting the definition of τ. Therefore,

$$\lim_{t\to\tau^-} z(t) = 0.$$

From this and (2.4) we conclude that $u'(\xi(\alpha),\alpha) = 0$.

(ii) We now consider the case $\tau_0 < \tau$, and again we suppose that $\tau = \infty$. At $t = \tau_0$ we have

$$z(\tau_0) > 0, \quad z'(\tau_0) = 0, \quad \text{and} \quad z''(\tau_0) \geq 0.$$

We again distinguish two cases:

$$\text{(ii}^*) \ \ 0 \leq \tau_0 < 1 \quad \text{and} \quad \text{(ii}^{**}) \ \ \tau_0 \geq 1.$$

(ii*) We claim that

$$z(t) > f(t) \quad \text{and} \quad z'(t) > 0 \quad \text{for } \tau_0 < t < \tau_0 + \varepsilon, \tag{2.11}$$

where ε is some small positive constant. If $z''(\tau_0) > 0$, this follows immediately from (2.6a). If $z''(\tau_0) = 0$, differentiation of (2.6a) yields $z'''(\tau_0) > 0$ if $\tau_0 > 0$ and $z'''(\tau_0) = 0$ if $\tau_0 = 0$. In the latter case one further differentiation of (2.6a) shows that $z^{iv}(\tau_0) > 0$. Thus, in all cases (2.11) holds.

This enables us to define

$$\tau_1 = \sup\{t > \tau_0 : z' > 0 \ \text{ on } \ (\tau_0, t)\}.$$

We shall show that

$$\tau_1 < \infty, \quad z'(\tau_1) = 0, \quad \text{and} \quad z''(\tau_1) < 0. \tag{2.12}$$

Suppose, to the contrary, that $\tau_1 = \infty$. Then, since $f(t) > 0$ for $t > \tau_+(\mu) = \sqrt{1 + \sqrt{\mu}}$, it follows from (2.10) that

$$(z^{3/4})'' < \frac{3}{4\gamma} z^{-1/4} \quad \text{for } t > \tau_+$$

or

$$y'' < \frac{3}{4\gamma} y^{-1/3} \quad \text{for } t > \tau_+, \tag{2.13}$$

where we have set $y = z^{3/4}$. If we now multiply (2.13) by $2y'$ and integrate over (τ_+, t), we find that

$$y' < \frac{3}{2\sqrt{\gamma}} \sqrt{y^{2/3} + C} \quad \text{for } t > \tau_+,$$

where C is a positive constant. Writing $w = y^{2/3} = z^{1/2}$, this inequality translates into

$$w' < \frac{1}{\sqrt{\gamma}} \sqrt{\frac{w + C}{w}} \quad \text{for } t > \tau_+.$$

Thus, since w is increasing, w' is uniformly bounded on (τ_+, ∞), so that

$$z(t) < A(1 + t)^2 \quad \text{for } t > 0 \tag{2.14}$$

for some positive constant A.

Remembering that $f(t) \sim \frac{1}{2} t^4$ as $t \to \infty$, it follows from (2.14) that $z(t) - f(t) \sim -\frac{1}{2} t^4$ as $t \to \infty$. Hence, since z is increasing, there exists a constant $K > 0$ such that

$$(z^{3/4})'' < -K(1 + t)^{3/2} \quad \text{for } t > t_1,$$

where t_1 is some sufficiently large number. Two integrations show that z cannot keep increasing indefinitely, contradicting our assumption that $\tau_1 = \infty$. Thus,

$$\tau_1 < \infty \quad \text{and} \quad z'(\tau_1) = 0.$$

To complete the proof of (2.12), in this case, we shall show that

$$\tau_1 > 1 \quad \text{and} \quad z''(\tau_1) < 0. \tag{2.15}$$

By (2.11), we can define

$$\tau^* = \sup\{t > \tau_0 : z > f \ \text{ on } \ (\tau_0, t)\}.$$

It then follows from (2.6a) that $z'' > 0$ on $(\tau_0, \tau^*]$. This implies that $\tau_1 > \tau^*$. Also, since $f' < 0$ on (0,1), it must be that $\tau_1 > 1$.

To prove the second assertion in (2.15), we need to show that $z''(\tau_1) < 0$. Suppose, to the contrary, that $z''(\tau_1) = 0$. Then, since $f' > 0$ on $(1, \infty)$, it follows that $z'''(\tau_1) < 0$, which implies that $z'' > 0$ and $z' < 0$ on a left-neighborhood of τ_1, contradicting the definition of τ_1. This completes the proof of (2.15).

It now follows from (2.6a) that $z'' < 0$ and $z' < 0$ for $t > \tau_1$ until $z = 0$ at some finite value τ. This contradicts our assumption that $\tau = \infty$. As in case (i) above it follows that $u'(\xi(\alpha), \alpha) = 0$.

(ii**) We now assume that $\tau_0 \geq 1$. The definition of τ_0 implies that $z''(\tau_0) \geq 0$. If $z''(\tau_0) > 0$, then $z' > 0$ on a right-neighborhood of τ_0 and we proceed as in the previous case. If $z''(\tau_0) = 0$, then $z'''(\tau_0) < 0$ if $\tau_0 > 1$, and it follows from (2.10) that $z'' < 0$ and $z' < 0$ to the right of τ_0 until $z = 0$ at a finite τ, a contradiction. If $\tau_0 = 1$ in this case, then $z'''(\tau_0) = 0$ as well, but $z^{iv}(\tau_0) < 0$. Again we find that $z'' < 0$ and $z' < 0$ to the right of τ_0 until $z = 0$ at a finite τ, a contradiction.

This completes the proof of part (a).

(b) Suppose, to the contrary, that there exist constants $\mu > 0$, $\alpha > 0$, and $\gamma > 0$ such that $\xi(\alpha) = \infty$. Then $u'(x) > 0$ for all $x > 0$ and by part (a), u is uniformly bounded on $\mathbf{R}^+$. Hence,

$$\lim_{x \to \infty} u(x, \alpha) \text{ exists} \overset{\text{def}}{=} \ell.$$

We distinguish three cases:

$$\text{(i)} \quad \ell > 1, \quad \text{(ii)} \quad 0 < \ell < 1, \quad \text{and} \quad \text{(iii)} \quad \ell = 1.$$

(i) Since $\ell(1 - \ell^2) < 0$ if $\ell > 1$, there exist a point $x_1 > 0$ and a constant $M > 0$ such that

$$\gamma u^{iv} - u'' < -M \quad \text{for } x > x_1.$$

When we integrate this inequality twice over (x_1, x), we obtain in turn

$$\gamma u'''(x) - u'(x) < A - Mx,$$
$$\gamma u''(x) - u(x) < Ax + B - \frac{1}{2}Mx^2,$$

where A and B are appropriate constants. Plainly, it is possible to choose $x_2 > x_1$ so large that

$$Ax + B < \frac{1}{4}Mx^2 \quad \text{for } x > x_2$$

and hence,

$$\gamma u''(x) < u(x) - \frac{1}{4}Mx^2 \quad \text{for } x > x_2.$$

Because u is uniformly bounded, this means that there exists a point $x_3 > x_2$ such that

$$\gamma u''(x) < -\frac{1}{8}Mx^2 \quad \text{for } x > x_3.$$

One more integration shows that $u'(x) \to -\infty$ as $x \to \infty$, which contradicts the assumption that $\xi(\alpha) = \infty$.

(ii) Since $\ell(1-\ell^2) > 0$ if $0 < \ell < 1$, there exist a point $y_1 > 0$ and a constant $N > 0$ such that

$$\gamma u^{iv} - u'' > N \quad \text{for } x > y_1.$$

Proceeding as in the previous case, we can find a point $y_2 > y_1$ such that

$$\gamma u''(x) > +\frac{1}{8}Nx^2 \quad \text{for } x > y_2,$$

and we conclude that $u'(x) \to \infty$ as $x \to \infty$, which contradicts the fact that u is uniformly bounded.

(iii) If $\ell = 1$, we conclude from (1.7) that there exists a point $x^* > 0$ such that

$$u'(x)u'''(x) \geq \frac{\mu}{8\gamma} > 0 \quad \text{for } x > x^*. \tag{2.16}$$

Hence, u'' has one sign on (x^*, ∞), so $\lim_{x\to\infty} u'(x) = m$ exists and $m \geq 0$. If $m > 0$, then $u(x) \to \infty$ as $x \to \infty$, violating the boundedness of u. Hence $m = 0$, so that by (2.16), $u'''(x) \to \infty$ as $x \to \infty$. Thus, $u''(x) \to \infty$, $u'(x) \to \infty$, and $u(x) \to \infty$ as $x \to \infty$, contradicting again the boundedness of u.

(c) Suppose, to the contrary, that there exist constants $\gamma > 0$ and $\alpha > 0$ such that $\xi(\alpha) = \infty$. Then $u'(x) > 0$ for all $x > 0$ and it follows as in part (b) that

$$\lim_{x\to\infty} u(x,\alpha) = 1. \tag{2.17}$$

Hence, $0 < u < 1$ for all $x > 0$, and we deduce from the differential equation that

$$\gamma u^{iv} - u'' > 0 \quad \text{for } x > 0. \tag{2.18}$$

By the maximum principle, u'' can have only one sign on $\mathbf{R}^+$, and therefore u' tends to a limit: $\lim_{x\to\infty} u'(x) = \ell_1 \geq 0$. If $\ell_1 > 0$, the solution will be unbounded on $\mathbf{R}^+$, so

$$\lim_{x\to\infty} u'(x) = 0. \tag{2.19}$$

In view of (2.18), $\lim_{x\to\infty}(\gamma u''' - u') = \ell_2$ exists, and therefore,

$$\lim_{x\to\infty} u'''(x) = \frac{\ell_2}{\gamma}.$$

Reasoning as before we find that

$$\lim_{x\to\infty} u'''(x) = 0. \tag{2.20}$$

Thus, since $\mu = 0$, it follows from (2.17)–(2.20) and the energy identity (1.7) that

$$\lim_{x\to\infty} (u, u', u'', u''') = (1,0,0,0) \qquad \text{and} \qquad u' > 0 \quad \text{on } [0,\infty). \tag{2.21}$$

If $\gamma \in (0, \frac{1}{8}]$, then (2.21) implies that $\alpha = \alpha_0$, where α_0 is the unique positive value of α for which a monotone, antisymmetric kink U exists. But this value has been excluded in the hypotheses.

If $\gamma > \frac{1}{8}$, then a linearization of (2.6a) around $u = 1$ shows that all four eigenvalues are complex with nonzero real and imaginary parts. Therefore, u cannot approach $u = 1$ monotonically, as asserted in (2.21). Thus, we have arrived at the final contradiction and Lemma 2.1 is proved.

We conclude this section with a lemma which restricts the admissible values of μ and α.

LEMMA 2.2. *Let $\gamma > 0$ and suppose that either*

$$\text{(a)} \qquad 0 \le \mu < 1 \quad \textit{and} \quad \alpha \ge \sqrt{\frac{1-\mu}{2}}$$

or

$$\text{(b)} \qquad \mu \ge 1 \quad \textit{and} \quad \alpha \ge 0.$$

Then the solution $u(\cdot, \alpha)$ of problem (2.1), (2.2) *cannot satisfy* (2.3).

Proof. Suppose, to the contrary, that there exist values of μ and α for which either (a) or (b) holds and which are such that the corresponding solution $u(x, \alpha)$ of problem (2.1), (2.2) does satisfy (2.3). We consider the two cases in succession.

(a) Observe that

$$z''(0, \alpha) = \frac{1}{\gamma\alpha^2}\Big(\alpha^2 - \frac{1-\mu}{2}\Big).$$

Thus, if $\alpha > \sqrt{\frac{1}{2}(1-\mu)}$, then $z''(0, \alpha) > 0$, and if $\alpha = \sqrt{\frac{1}{2}(1-\mu)}$, then $z''(0, \alpha) = 0$. However, in the second case $z'''(0, \alpha) > 0$. Therefore, in both cases $z'' > 0$ on an interval $(0, \delta)$.

It follows by the arguments used in the proof of part (a) of Lemma 2.1 that $z''(t) > 0$ and $z'(t) > 0$ for all $t \in (0, \tau_+(\mu)]$. In terms of the function $u(\cdot, \alpha)$ this means that there exists a point x_1 such that

$$u' > 0, \quad u'' > 0, \quad \text{and} \quad u''' > 0 \quad \text{on } (0, x_1) \qquad \text{and} \qquad u(x_1) > 1.$$

For condition (2.3b) to hold, there must be a first $x_2 > x_1$ such that

$$u''(x_2) = 0 \quad \text{and} \quad u'''(x_2) \le 0.$$

Thus, $u'' > 0$ and, hence, also $u' > 0$ on $(0, x_2)$, so that $\xi(\alpha) > x_2$. By equation (2.1a), we have $u^{iv}(x_2) < 0$, so

$$u''' < 0 \quad \text{and} \quad u'' < 0 \quad \text{on } (x_2, \xi(\alpha)].$$

In particular, $u'''(\xi(\alpha), \alpha) < 0$, which contradicts (2.3b).

(b) In this case, if $\alpha > 0$ then $z''(0, \alpha) > 0$ and we can proceed using the same arguments we used in part (a) to complete the proof. We omit the details for the sake of brevity.

If $\alpha = 0$, then by (1.7) we must have $\mu = 1$, so that $u(0) = 0$, $u'(0) = 0$, $u''(0) = 0$. If $u'''(0) = 0$ as well, then u is the trivial solution, so we must assume that $u'''(0) \neq 0$. Without loss of generality we may assume that $u'''(0) > 0$. Then $u' > 0$ and $u'' > 0$ in a right-neighborhood of the origin; hence, $z(t) > 0$ and $z'(t) > 0$ for small values of t. We may now proceed as in part (a) to show that u cannot be a periodic solution which satisfies (2.3). Theorem B is an immediate consequence of part (b) of Lemma 2.2.

3. Existence and uniqueness of periodic solutions: $0 < \mu < 1,\ \gamma > 0$. In this section we focus our attention on the parameter range $0 < \mu < 1,\ \gamma > 0$. We prove Theorem A and a uniqueness theorem for a more restricted range of values of μ, i.e., $\mu \in (0, \frac{4}{9}]$.

THEOREM 3.1. *Let $\mu \in (0,1)$ and $\gamma > 0$. Then there exists a periodic solution $u(x)$ such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} < 1.$$

The proof proceeds via a sequence of lemmas.

We define the shooting set

$$\mathcal{S} = \{\hat{\alpha} > 0 : u(\xi(\alpha),\alpha) < 1,\ u''(\xi(\alpha),\alpha) < 0, \text{ and } u'''(\xi(\alpha),\alpha) < 0 \text{ for } 0 < \alpha < \hat{\alpha}\}.$$

LEMMA 3.2. *We have*

(a) $\xi \in C^1(\mathcal{S})$.

(b) *$\mathcal{S}$ is an open interval.*

(c) $$u(\xi(\alpha),\alpha) < \sqrt{1-\sqrt{\mu}} \quad \textit{if } \alpha \in \mathcal{S}.$$

Proof. (a) Let $\alpha \in \mathcal{S}$. Then, at $\xi = \xi(\alpha)$ we have

$$u'(\xi(\alpha),\alpha) = 0 \quad \text{and} \quad u''(\xi(\alpha),\alpha) < 0.$$

Hence, by the implicit function theorem, $\xi \in C^1(\mathcal{S})$.

(b) Since the inequalities in the definition of $\mathcal{S}$ are strict, the assertion follows immediately from part (a) and the continuous dependence of solutions on initial data.

Part (c) follows at once from the energy identity (1.7).

In the following lemma we show that $\mathcal{S}$ is nonempty. Define

$$\overline{\alpha} = \min\left\{\frac{\sqrt{1-\mu}}{2}, \sqrt{\frac{3(1-\mu)}{24\gamma+7}}\right\}.$$

LEMMA 3.3. $(0,\overline{\alpha}) \subset \mathcal{S}$.

Proof. Let $\alpha \in (0,\overline{\alpha})$. Observe that (1.7) implies that $u'''(0) < 0$, since $0 < \mu < 1$ and $\alpha < \overline{\alpha}$. As we increase x and as long as $u''' < 0$, it follows that $u'' < 0$, $u' < \alpha$, and $u(x) < \alpha x$. Thus, as long as $u \geq 0$ and $u''' < 0$, it follows from (2.1a) that

$$u^{iv}(x) < \frac{\alpha}{\gamma}x. \tag{3.1a}$$

We integrate this inequality three times to obtain

$$u'''(x) < \beta + \frac{\alpha}{2\gamma}x^2, \tag{3.1b}$$

$$u''(x) < \beta x + \frac{\alpha}{6\gamma}x^3, \tag{3.1c}$$

$$u'(x) < \alpha + \frac{\beta}{2}x^2 + \frac{\alpha}{24\gamma}x^4. \tag{3.1d}$$

Since we assume that $\alpha < \frac{1}{2}\sqrt{1-\mu}$, the energy identity (1.7) implies that

$$\beta < -\frac{1-\mu}{8\gamma\alpha}.$$

Hence, the right-hand sides of (3.1b) and (3.1c) are negative for $0 < x \leq 1$ and, since $\alpha < 1$, it follows that $u < 1$ on $[0, 1]$ as long as $u' \geq 0$.

On the other hand, because we have chosen $\alpha < \sqrt{\frac{3(1-\mu)}{24\gamma+7}}$, the right-hand side of (3.1d) is negative at $x = 1$. Thus, there must exist a first zero ξ of u' on $(0, 1)$, where $u < 1$, $u'' < 0$, and $u''' < 0$, so that $\alpha \in \mathcal{S}$.

Define

$$\alpha^* = \sup \mathcal{S}.$$

Lemma 3.4. *We have*

$$\alpha^* \leq \sqrt{\frac{1-\mu}{2}}.$$

Proof. It follows from (1.7) that

$$\begin{aligned} 2\gamma u' u''' &\geq (u')^2 - \frac{1}{2}\{(1-u^2)^2 - \mu\} \\ &> (u')^2 - \frac{1-\mu}{2} \quad \text{for } 0 < u < 1. \end{aligned}$$

Thus, if $\alpha^2 > (1-\mu)/2$, then $u''' > 0$, $u'' > 0$, and $u' > \alpha > 0$ as long as $0 < u \leq 1$, so that u' cannot have a first zero ξ such that $u(\xi) < 1$.

Lemma 3.5. *We have*

$$0 < u(\xi(\alpha^*), \alpha^*) < 1, \quad u''(\xi(\alpha^*), \alpha^*) < 0, \quad \textit{and} \quad u'''(\xi(\alpha^*), \alpha^*) = 0.$$

Proof. Suppose that $u(\xi(\alpha^*), \alpha^*) \geq 1$. Then by the continuous dependence of $u(\cdot, \alpha)$ on α on compact intervals, it follows that $u(\xi(\alpha), \alpha) > \sqrt{1 - \sqrt{\mu}}$ for all α in a small enough neighborhood of α^*. Since $(0, \alpha^*) \subset \mathcal{S}$, this contradicts Lemma 3.2 and we conclude that

$$u(\xi(\alpha^*), \alpha^*) < 1. \tag{3.2}$$

Thus, in the limit as α increases toward α^*, the first inequality in the definition of $\mathcal{S}$ continues to hold, and we wish to prove that the second one continues to hold as well. Suppose that the second inequality fails. It follows from the definition of ξ that $u''(\xi(\alpha^*), \alpha^*) \leq 0$. Hence, we suppose that

$$u''(\xi(\alpha^*), \alpha^*) = 0. \tag{3.3}$$

In what follows we shall write $\xi^* = \xi(\alpha^*)$ and $u^* = u(\xi^*, \alpha^*)$.

To show that (3.3) leads to a contradiction we proceed via a series of steps.

Step 1. We show that (3.3) implies that

$$u'''(\xi^*, \alpha^*) > 0. \tag{3.4}$$

Suppose that $u'''(\xi^*, \alpha^*) < 0$. Then $u'' > 0$ and $u' < 0$ in a left-neighborhood of ξ^*, contradicting the definition of ξ^*.

Next, suppose that $u'''(\xi^*, \alpha^*) = 0$. Then, since $u^* \in (0, 1)$ by (3.2), it follows from the differential equation that $u^{iv}(\xi^*, \alpha^*) > 0$, so that $u''' < 0$, $u'' > 0$, and $u' < 0$

in a left-neighborhood of ξ^*. This means that u' has a zero on $(0,\xi^*)$ contradicting again the definition of ξ^*.

Thus, if (3.3) holds, then so does (3.4).

Step 2. We show that

$$\xi(\alpha) \to \xi(\alpha^*) \quad \text{as } \alpha \to \alpha^*,\ \alpha \in \mathcal{S}. \tag{3.5}$$

First, it follows from (3.3), (3.4), and (2.1a) that

$$u'''(x,\alpha^*) > 0, \quad u''(x,\alpha^*) > 0, \quad \text{and} \quad u'(x,\alpha^*) > 0$$

for $x > \xi^*$ until $u(x_0,\alpha^*) = 1$ at a finite $x_0 > \xi(\alpha^*)$.

Next, let $\varepsilon > 0$ be small and arbitrarily chosen to satisfy

$$0 < \xi(\alpha^*) - \varepsilon < \xi(\alpha^*) < \xi(\alpha^*) + \varepsilon < x_0.$$

Since $[0,x_0]$ is compact, it follows from the definition of α^*, part (b) of Lemma 3.2, and continuity that there exists a $\delta = \delta(\varepsilon) > 0$ such that if $0 < \alpha^* - \alpha < \delta$, then $\alpha \in \mathcal{S}$,

$$u(x_0,\alpha) > \sqrt{1-\sqrt{\mu}}, \tag{3.6a}$$

and

$$u'(x_0,\alpha) > 0 \quad \text{for all } x \in [0,\xi(\alpha^*)-\varepsilon] \cup [\xi(\alpha^*)+\varepsilon, x_0]. \tag{3.6b}$$

From (3.6a), part (c) of Lemma 3.2, (3.6b), and the definition of $\xi(\alpha)$ we conclude that $\xi(\alpha) \in (\xi(\alpha^*)-\varepsilon, \xi(\alpha^*)+\varepsilon)$ if $0 < \alpha^* - \alpha < \delta$. This implies (3.5).

Step 3. The contradiction. It follows from (3.5) that

$$u'''(\xi(\alpha),\alpha) \to u'''(\xi(\alpha^*),\alpha^*) \quad \text{as } \alpha \to \alpha^*,\ \alpha \in \mathcal{S}.$$

Because $u'''(\xi(\alpha),\alpha) < 0$ for all $\alpha \in \mathcal{S}$, this implies that

$$u'''(\xi(\alpha^*),\alpha^*) \le 0,$$

which contradicts (3.4) and we conclude that (3.3) cannot hold. Thus,

$$u''(\xi(\alpha^*),\alpha^*) < 0. \tag{3.7}$$

To complete the proof we suppose that

$$u'''(\xi(\alpha^*),\alpha^*) < 0.$$

Then $\alpha^* \in \mathcal{S}$ and since $\mathcal{S}$ is open, α^* cannot be the supremum of $\mathcal{S}$. Therefore,

$$u'''(\xi(\alpha^*),\alpha^*) = 0$$

and the lemma is proved.

Corollary 3.6. *We have*

$$u(\xi(\alpha^*),\alpha^*) < \sqrt{1-\sqrt{\mu}} \quad \textit{and} \quad \alpha^* < \sqrt{\frac{1-\mu}{2}}.$$

Proof. The first inequality follows as in Lemma 3.2(c) from the energy identity (1.7). However, because we know from Lemma 3.5 that $u''(\xi^*, \alpha^*) < 0$, we now obtain strict inequality.

The second inequality is proved if we can rule out equality from Lemma 3.4. Thus, suppose that $(\alpha^*)^2 = (1-\mu)/2$. Then $u^{(i)}(0) = 0$ for $i = 2, 3, 4$ and $u^{(5)}(0) > 0$. Hence, since $u^{(i)} > 0,\ i = 2, 3, 4$ in a right-neighborhood of the origin, we conclude that $u' > 0$ as long as $u \le 1$, so that $u(\cdot, \alpha^*)$ cannot yield a periodic solution such that $u(\xi^*, \alpha^*) < 1$.

Proof of Theorem 3.1. It follows from Lemma 3.5 that the solution $u(x, \alpha^*)$ of problem (2.1) satisfies the conditions (2.3) at $\xi = \xi(\alpha^*)$ and so can be continued to yield a periodic solution with period $4\xi(\alpha^*)$.

Concerning *uniqueness*, we give the following partial result.

Lemma 3.7. *Let $\gamma > 0$ and $\frac{4}{9} \le \mu < 1$. Then there exists a unique periodic solution u which satisfies* (2.1), (2.2) *and is such that* $\max|u| < 1$.

Proof. Suppose that there are values $\gamma > 0$ and $\mu \in [\frac{4}{9}, 1)$ such that there exist two distinct periodic solutions u_1 and u_2 with $\max|u_i| < 1$ $(i = 1, 2)$. Let α_1 and α_2 be their respective slopes at $x = 0$. Since

$$\frac{d\beta}{d\alpha} = \frac{1}{2\gamma} + \frac{1-\mu}{4\gamma\alpha^2} > 0,$$

it follows that

$$\alpha_1 < \alpha_2 \quad \Rightarrow \quad u_1'''(0) < u_1'''(0) < 0.$$

Let $w = u_1 - u_2$. Then, by the mean value theorem,

$$\begin{cases} \gamma w^{iv} = w'' + (1 - 3\tilde{u}^2)w, & \text{(3.8a)} \\ w(0) = 0, \quad w'(0) < 0, \quad w''(0) = 0, \quad \text{and} \quad w'''(0) < 0, & \text{(3.8b)} \end{cases}$$

where $\tilde{u}$ is a function whose values lie between those of u_1 and u_2. Since $\frac{4}{9} \le \mu < 1$ and $\max|u_i| < \sqrt{1 - \sqrt{\mu}}$, it follows that

$$|\tilde{u}(x)| \le \frac{1}{\sqrt{3}} \quad \text{for } x \in \mathbf{R}$$

and, therefore,

$$1 - 3\tilde{u}^2(x) \ge 0 \quad \text{for } x \in \mathbf{R}.$$

Thus, we conclude from (3.8a) and (3.8b) that

$$w^{iv} < 0,\ w''' < 0,\ w'' < 0, \ \text{ and } \ w' < 0 \quad \text{for } x \in \mathbf{R}.$$

This implies that $w(x) \to -\infty$ as $x \to \infty$. Because $|w(x)| \le |u_1(x)| + |u_2(x)| \le \frac{2}{\sqrt{3}}$ for all $x \in \mathbf{R}$, this is not possible and we have a contradiction.

We conjecture that for every $\gamma > 0$ and $0 < \mu < 1$, there exists a unique periodic solution with maximum less than 1.

4. Nonexistence of periodic solutions. In this section we restrict our attention to the parameter regime

$$0 \leq \mu < 1 \quad \text{and} \quad \gamma > 0. \tag{4.1}$$

In Lemma 2.2 we already proved that no periodic solution can exist if $\mu \geq 1$ and $\gamma > 0$. We wish to determine the largest possible range of values of μ and γ in this regime in which no periodic solution exists. We emphasize again that by a periodic solution we mean a solution of problem (2.1)–(2.2) which satisfies (2.3).

In the analysis below we extend Lemma 2.2 and prove two nonexistence theorems for $0 < \gamma \leq \frac{1}{8}$. In the previous section we found that there exist periodic solutions for every $\gamma > 0$ when $0 < \mu < 1$ and that their maxima lie *below* $u = 1$. In our first nonexistence theorem we set $\mu = 0$ and show that such solutions cease to exist when $0 < \gamma \leq \frac{1}{8}$. This range of γ-values is optimal; in the next section we shall show that when $\mu = 0$ and $\gamma > \frac{1}{8}$ such periodic solutions do exist.

In section 5 it is shown that when $\mu = 0$ and $\gamma > \frac{1}{8}$, there exist periodic solutions with maxima *above* $u = 1$. In our second nonexistence theorem we shall show that when $\mu = 0$, this range of γ-values is also optimal and that no periodic solutions with maxima above $u = 1$ exist when $\gamma \leq \frac{1}{8}$.

Before proving these two theorems, we establish three technical lemmas. For this we recall that if $u(\cdot, \alpha)$ is a periodic solution, then the corresponding solution $z(\cdot, \alpha)$ of problem (2.6) has the properties

$$\lim_{t \to \tau(\alpha)^-} z(t, \alpha) = 0 \quad \text{and} \quad \lim_{t \to \tau(\alpha)^-} \sqrt{z(t, \alpha)} z''(t, \alpha) = 0, \tag{2.8b}$$

where $\tau(\alpha)$ is finite and defined by

$$\tau(\alpha) = \sup\{t > 0 : z(\cdot, \alpha) > 0 \;\text{ on }\; [0, t)\}.$$

LEMMA 4.1. *Let $0 \leq \mu < 1$ and $\gamma > 0$. If z corresponds to a periodic solution, then*

(a)
$$z(0) < \frac{1 - \mu}{2},$$

(b)
$$z'(t) < 0 \quad \textit{for } 0 < t < \tau.$$

Proof. (a) Since $z(0) = \alpha^2$, the assertion follows at once from Lemma 2.2.

(b) It follows from part (a) and (2.6a) that $z''(0) < 0$. Hence, $z' < 0$ in an interval $(0, \varepsilon)$ for some small $\varepsilon > 0$. If z' vanishes at a first $\tau_0 \in (0, \tau)$, then

$$z(\tau_0) > 0, \quad z'(\tau_0) = 0, \quad \text{and} \quad z''(\tau_0) \geq 0. \tag{4.2}$$

If $z''(\tau_0) = 0$, then $z(\tau_0) = f_\mu(\tau_0)$ by (2.6a) and a differentiation of (2.6a) yields

$$z'''(\tau_0) = -\frac{f'_\mu(\tau_0)}{\gamma z(\tau_0)} = -\frac{2\tau_0(\tau_0^2 - 1)}{\gamma\, z(\tau_0)}. \tag{4.3}$$

We shall discuss the cases (i) $\tau_0 = 1$, (ii) $\tau_0 < 1$, and (iii) $\tau_0 > 1$ in succession.

(i) $\tau_0 = 1$. It follows from (2.6a) that

$$z(\tau_0) = f_\mu(\tau_0) = -\frac{1}{2}\mu \leq 0,$$

which contradicts (4.2).

(ii) $\tau_0 < 1$. It follows from (4.3) that $z'''(\tau_0) > 0$, so that $z'' < 0$ and $z' > 0$ in a left-neighborhood $(\tau_0 - \varepsilon, \tau_0)$ of τ_0. But this contradicts the definition of τ_0.

(iii) $\tau_0 > 1$. It follows from (4.3) that $z'''(\tau_0) < 0$. Therefore, $z'' < 0$ and $z' < 0$ in a small interval $(\tau_0, \tau_0 + \varepsilon)$. When we differentiate (2.6a) and divide by $\sqrt{z}$, we obtain

$$(\sqrt{z}\, z'')' = \frac{z' - f'_\mu}{\gamma\sqrt{z}}. \tag{4.4}$$

Since $\tau_0 > 1$ and $f'_\mu > 0$ on $(1, \infty)$, it follows from (4.4) that $z'' < 0$ and $z' < 0$ on the entire interval (τ_0, τ). Integration of (4.4) over (τ_0, τ) shows that

$$\lim_{t \to \tau^-} \sqrt{z(t)}\, z''(t) = \int_{\tau_0}^{\tau} \frac{z' - f'_\mu}{\gamma\sqrt{z}}\, ds < 0,$$

contradicting (2.8b).

Thus, because (i), (ii), and (iii) lead to contradictions, we must conclude that $z''(\tau_0) > 0$. This implies by (2.6a) that $z(\tau_0) > f_\mu(\tau_0)$. Hence, $z' > 0$ in a right-neighborhood of τ_0. Because $z(\tau) = 0$, there must exist a first $\tau_1 > \tau_0$ where $z'(\tau_1) = 0$. We assert that $\tau_1 > 1$. To see this, note that (2.6a) implies that $z'' > 0$; hence, $z' > 0$ as long as $z > f_\mu$. Because $f'_\mu < 0$ on $(0, 1)$, it follows that

$$z(t) > z(\tau_0) > f_\mu(\tau_0) > f_\mu(t) \quad \text{on } \tau_0 < t \leq 1.$$

This means that $z' > 0$ on $(\tau_0, 1]$ and, hence, $\tau_1 > 1$.

At $t = \tau_1$ we have $z''(\tau_1) \leq 0$. If $z''(\tau_1) < 0$, then $z'' < 0$ and $z' < 0$ on an interval $(\tau_1, \tau_1 + \varepsilon)$. Reasoning as before, using (4.4) again, we find that $z' < 0$ on the entire interval (τ_1, τ) and that $\lim_{t \to \tau^-} \sqrt{z(t)}\, z''(t) < 0$, which contradicts (2.8b). If $z''(\tau_1) = 0$, then, as in (4.3), we find that

$$z'''(\tau_1) = -\frac{2\tau_1(\tau_1^2 - 1)}{\gamma\, z(\tau_1)} < 0.$$

Therefore, $z'' < 0$ and $z' < 0$ on a right-neighborhood of τ_1 and we can repeat the previous argument to obtain a contradiction of (2.8b).

This leads us to the conclusion that $z' < 0$ on $(0, \tau)$ and the lemma is proved.

Lemma 4.2. *Suppose that z corresponds to a periodic solution and that $f_\mu(\tau) \neq 0$. Then*

$$z''(\tau) = \lim_{t \to \tau^-} z''(t) = \frac{2}{\gamma}\left\{1 + \frac{\sqrt{\gamma}}{2}\frac{f'_\mu(\tau)}{\sqrt{f_\mu(\tau)}}\right\}. \tag{4.5}$$

Proof. Because u' and u''' both vanish as $x \to \xi^-$, or $t \to \tau^-$, it follows from l'Hôpital's rule that

$$z''(\tau) = 2\lim_{x \to \xi^-} \frac{u'''(x)}{u'(x)} = 2\lim_{x \to \xi^-} \frac{u^{iv}(x)}{u''(x)} = \frac{2}{\gamma}\left(1 + \frac{u - u^3}{u''}\right), \tag{4.6}$$

where the last term is evaluated at $x = \xi$. By the energy identity (1.7) we have

$$(u'')^2 = \frac{1}{\gamma} f_\mu(u) \quad \text{at } x = \xi,$$

so that

$$u'' = -\frac{1}{\sqrt{\gamma}}\sqrt{f_\mu(u)} \quad \text{at } x = \xi.$$

If we substitute this expression for u'' into (4.6) and remember that $u(\xi) = \tau$, the assertion follows.

Define

$$H = z\Big(z'' - \frac{1}{\gamma}\Big) - \frac{\mu}{2\gamma} \tag{4.7}$$

and

$$\tau_0 = \sup\{t \in (0, \tau) : z' < 0 \;\text{ on }\; (0, t)\}.$$

Lemma 4.3. *Let $0 \le \mu < 1$ and $0 < \gamma \le \frac{1}{8}$, and let z be the solution of problem* (2.6). *Then*

$$H(t) < 0 \quad \textit{for } 0 \le t < \tau^* = \min\{\tau_0, 1\}.$$

Proof. Observe that we can write H as

$$H = \frac{(z')^2}{4} - \frac{(t^2-1)^2}{2\gamma}. \tag{4.8}$$

Hence

$$H(0) = -\frac{1}{2\gamma} < 0$$

and it follows that $H < 0$ in a right-neighborhood of the origin. Suppose that H first vanishes at a point $t_0 \in (0, \tau^*)$. Then

$$H(t_0) = 0 \quad \text{and} \quad H'(t_0) \ge 0. \tag{4.9}$$

We deduce from (4.8), (4.9), and part (b) of Lemma 4.1 that, since $t_0 < 1$,

$$z'(t_0) = -\sqrt{\frac{2}{\gamma}}\,(1 - t_0^2). \tag{4.10}$$

For H' we differentiate (4.8) and use (4.10) to obtain

$$H'(t_0) = -(1 - t_0^2)\left\{\frac{z''}{\sqrt{2\gamma}} - \frac{2t_0}{\gamma}\right\}. \tag{4.11}$$

Since $\mu \ge 0$ and H vanishes at t_0, it follows from (4.7) and (4.9) that

$$z\Big(z'' - \frac{1}{\gamma}\Big) = \frac{\mu}{2\gamma} \ge 0 \quad \text{at } t = t_0.$$

Hence $z''(t_0) \ge \frac{1}{\gamma}$, and we conclude from (4.11) that

$$H'(t_0) \le -\frac{1}{\gamma}(1 - t_0^2)\left\{\frac{1}{\sqrt{2\gamma}} - 2t_0\right\} < 0$$

because $0 < t_0 < 1$ and $\gamma \le \frac{1}{8}$. This contradicts (4.9) and the lemma is proved.

We are now ready to prove the two nonexistence theorems for $0 < \gamma \le \frac{1}{8}$. As was explained earlier, we know that there exist periodic solutions for these values of γ when $0 < \mu < 1$ and that they do not exceed $u = 1$. In the first theorem we show that such periodic solutions no longer exist when $\mu = 0$. In the second theorem we show that if $\mu \in [0, 1)$, then there exist no periodic solutions which exceed $u = 1$.

THEOREM 4.4. *Let $\mu = 0$ and $0 < \gamma \le \frac{1}{8}$. Then there exists no periodic solution $u(x)$ such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} < 1.$$

Proof. Suppose, to the contrary, that there exists a periodic solution u whose maximum is less than 1. Let z correspond to u. Then $\tau < 1$, Lemma 4.1 implies that $z' < 0$ on $(0, \tau)$, and we deduce from Lemma 4.3 that $H < 0$ on $(0, \tau)$. Thus, by (4.7),

$$z''(t) < \frac{1}{\gamma} \quad \text{for } 0 < t < \tau \tag{4.12}$$

and, in particular,

$$z''(\tau) \le \frac{1}{\gamma}.$$

From this last inequality and Lemma 4.2, we conclude that

$$f'(\tau) \le -\frac{1}{\sqrt{\gamma}}\sqrt{f(\tau)}.$$

Therefore, because of (2.7),

$$\tau \ge \frac{1}{\sqrt{8\gamma}}.$$

Since $\gamma \le \frac{1}{8}$, this means that we must have $\tau \ge 1$, a contradiction.

THEOREM 4.5. *Let $0 \le \mu < 1$ and $0 < \gamma \le \frac{1}{8}$. Then there exists no periodic solution $u(x)$ such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} > 1.$$

Proof. Suppose that there exists a periodic solution u whose maximum is greater than 1. Let z correspond to u. Then $\tau > 1$. Since Lemma 4.1 implies that $z' < 0$ on $(0, \tau)$, it follows that $z'(1) < 0$.

To force a contradiction we shall show that $z'(1) > 0$. The proof of Lemma 3.4 shows that this is the case when $\alpha > \alpha_\mu = \sqrt{(1-\mu)/2}$, and by continuity this will remain so until $z'(1) = 0$ for some $\tilde{\alpha} < \alpha_\mu$. When $z'(1) = 0$, it follows from (4.8) that

$$H(1) = 0 \quad \text{and} \quad H'(1) = 0.$$

In addition,

$$H''(1) \ge \frac{1 - 8\gamma}{2\gamma^2} + \frac{\mu}{2\gamma^2 z},$$

where strict inequality holds if $\mu > 0$ and equality holds if $\mu = 0$. Thus,

$$H''(1) > 0 \quad \text{if } (\mu, \gamma) \neq \left(0, \frac{1}{8}\right),$$

in which case $H' < 0$ and $H > 0$ in a left-neighborhood of $t = 1$. Since $H < 0$ on $(0, 1)$ by Lemma 4.3, we have a contradiction.

On the other hand, if $\mu = 0$ and $\gamma = \frac{1}{8}$, then $H''(1) = 0$ and we have to consider higher derivatives. We find that

$$H'''(1) = -\frac{12}{\gamma} < 0.$$

Therefore, in this case, $H'' > 0$, $H' < 0$, and $H > 0$ in a left-neighborhood of $t = 1$ and by Lemma 4.3, we have once again arrived at a contradiction.

5. Existence of periodic solutions: $\mu = 0$, $\gamma > \frac{1}{8}$.

In the previous section we saw that if $\mu = 0$, then there are no periodic solutions for $0 < \gamma \leq \frac{1}{8}$. In this section we shall show that there do exist periodic solutions when $\gamma > \frac{1}{8}$, both with a maximum less than 1 and with a maximum greater than 1. The method of proof is similar to the one used in section 3.

THEOREM 5.1. *Let $\mu = 0$ and $\gamma > \frac{1}{8}$. Then there exists a periodic solution u such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} < 1.$$

We recall from Lemma 2.1 that if $\mu = 0$ and $\gamma > \frac{1}{8}$, then

$$\xi(\alpha) < \infty \quad \text{and} \quad u'(\xi(\alpha), \alpha) = 0 \text{ for every } \ \alpha > 0.$$

Continuing as in section 3, we set

$$\mathcal{S} = \{\hat{\alpha} > 0 : u(\xi(\alpha), \alpha) < 1, \ u''(\xi(\alpha), \alpha) < 0, \text{ and } u'''(\xi(\alpha), \alpha) < 0 \text{ for } 0 < \alpha < \hat{\alpha}\}.$$

Reproducing the proofs of Lemmas 3.2–3.4, we establish the following properties of ξ and $\mathcal{S}$.

LEMMA 5.2. *Let $\mu \in [0, 1)$ and $\gamma > 0$.*

(a) *$\xi \in C^1(\mathcal{S})$.*

(b) *The set $\mathcal{S}$ is a nonempty, open interval of the form $(0, \alpha^*)$.*

(c) $$\min\left\{\frac{\sqrt{1-\mu}}{2}, \sqrt{\frac{3(1-\mu)}{24\gamma + 7}}\right\} \leq \alpha^* \leq \sqrt{\frac{1-\mu}{2}}.$$

We must still determine the properties of $u(\cdot, \alpha^*)$. This will be done in the next lemma.

LEMMA 5.3. *Let $\mu = 0$ and $\gamma > 0$. Then*

$$u(\xi(\alpha^*), \alpha^*) < 1, \quad u''(\xi(\alpha^*), \alpha^*) < 0, \quad \text{and} \quad u'''(\xi(\alpha^*), \alpha^*) = 0.$$

Proof. We first show that $u''(\xi^*, \alpha^*) < 0$, where we have written $\xi^* = \xi(\alpha^*)$.

From the definition of ξ we conclude that $u''(\xi^*, \alpha^*) \leq 0$. We claim that $u''(\xi^*, \alpha^*) < 0$. Thus, suppose to the contrary that

$$u''(\xi^*, \alpha^*) = 0. \tag{5.1}$$

Then, by the energy identity (1.7),

$$u(\xi^*, \alpha^*) = 1.$$

We assert that (5.1) implies that

$$u'''(\xi^*, \alpha^*) > 0. \tag{5.2}$$

Suppose that $u'''(\xi^*, \alpha^*) < 0$. Then $u'' > 0$ and $u' < 0$ in a left-neighborhood of ξ^*, contradicting the definition of ξ^*. On the other hand, if $u'''(\xi^*, \alpha^*) = 0$, then by uniqueness, $u(x) = 1$ for all $x \in \mathbf{R}$, a contradiction. Thus, indeed, (5.2) holds.

As in the proof of Lemma 3.5 it follows that

$$\xi(\alpha) \to \xi(\alpha^*) \quad \text{as } \alpha \to \alpha^*,\ \alpha \in \mathcal{S},$$

which means that

$$u'''(\xi(\alpha), \alpha) \to u'''(\xi(\alpha^*), \alpha^*) \quad \text{as } \alpha \to \alpha^*,\ \alpha \in \mathcal{S}.$$

Because $u'''(\xi(\alpha), \alpha) < 0$ for all $\alpha \in \mathcal{S}$, we arrive in the limit as $\alpha \to \alpha^*$ at

$$u'''(\xi(\alpha^*), \alpha^*) \leq 0, \tag{5.3}$$

which contradicts (5.2). Thus, (5.1) must be false; hence,

$$u''(\xi(\alpha^*), \alpha^*) < 0. \tag{5.4}$$

It follows from (5.4) and the energy identity (1.7) that

$$u^* = u(\xi(\alpha^*), \alpha^*) \neq 1$$

and that $\xi(\alpha)$ is continuous at $\alpha = \alpha^*$. Therefore, by the continuous dependence on initial data, if $u^* > 1$, then $u(\xi(\alpha), \alpha) > 1$ for α in a left-neighborhood of α^*. Since $(0, \alpha^*) \subset \mathcal{S}$, the definition of $\mathcal{S}$ shows that this is impossible. Thus,

$$u(\xi(\alpha^*), \alpha^*) < 1. \tag{5.5}$$

Finally, regarding u''', we must have equality in (5.3). For if

$$u'''(\xi(\alpha^*), \alpha^*) < 0,$$

then continuity implies that $\alpha^* < \sup \mathcal{S}$, a contradiction.

Thus, we have shown that the solution $u(x, \alpha^*)$ of problem (2.1) satisfies the properties (2.3) at $x = \xi(\alpha^*)$ and this yields a periodic solution of which, by (5.5), the maximum is less than 1. This completes the proof of Theorem 5.1.

In the next theorem we find periodic solutions whose maxima exceed unity.

THEOREM 5.4. *Let $\mu = 0$ and $\gamma > \frac{1}{8}$. Then there exists a periodic solution u such that*

$$\max\{|u(x)| : x \in \mathbf{R}\} > 1.$$

We now take the shooting set from those values of α for which the maximum of u *exceeds* 1. Specifically, we define

$$\mathcal{T} = \{\hat{\alpha} > 0 : u(\xi(\alpha), \alpha) > 1,\ u''(\xi(\alpha), \alpha) < 0,\ \text{and } u'''(\xi(\alpha), \alpha) < 0 \text{ for } \alpha > \hat{\alpha}\}.$$

Lemma 5.5. *We have*

$$\left(\frac{1}{\sqrt{2}},\infty\right)\subset\mathcal{T}.$$

Proof. It follows from (1.7) that

$$2\gamma u'u'''\geq (u')^2-\frac{1}{2}(1-u^2)^2.$$

Thus, if $\alpha^2>\frac{1}{2}$, then $u'''>0$, $u''>0$, and $u'>0$ as long as $0<u\leq 1$. Hence, u first reaches 1 at a finite value x_1, where

$$u'(x_1)>0,\quad u''(x_1)>0,\quad u'''(x_1)>0.$$

Therefore, at $x=\xi$ we have

$$u(\xi)>1,\quad u'(\xi)=0,\quad u''(\xi)<0, \tag{5.6}$$

where the last inequality is strict because of the energy identity (1.7). Hence, u'' has a first zero at a point $x_2\in(x_1,\xi)$. At this point we have

$$u(x_2)>1,\quad u''(x_2)=0,\quad u'''(x_2)\leq 0.$$

Since, by equation (2.1a), $u^{iv}<0$ when both $u>1$ and $u''\leq 0$, it follows that

$$u'''(\xi)<0. \tag{5.7}$$

From (5.6) and (5.7) we deduce that for any $\alpha>\frac{1}{\sqrt{2}}$,

$$u(\xi(\alpha),\alpha)>1,\quad u''(\xi(\alpha),\alpha)<1,\quad\text{and}\quad u'''(\xi(\alpha),\alpha)<0,$$

so that $(\frac{1}{\sqrt{2}},\infty)\subset\mathcal{T}$.

As with Lemma 5.2, we can prove the following properties of ξ.

Lemma 5.6. *We have*

(a) $\xi\in C^1(\mathcal{T})$,

(b) *the set $\mathcal{T}$ is an open interval of the form* (α_*,∞),

(c) $$\alpha_*\geq\alpha^*,$$

where $\alpha^*=\sup\mathcal{S}$ as defined in Lemma 5.2, *part* (b).

In the next lemma we list again the important properties of $u(\xi(\alpha),\alpha)$ at $\alpha=\alpha_*$.

Lemma 5.7. *We have*

$$u(\xi_*,\alpha_*)>1,\quad u''(\xi_*,\alpha_*)<0,\quad\textit{and}\quad u'''(\xi_*,\alpha_*)=0,$$

where we have written $\xi_*=\xi(\alpha_*)$.

Proof. From the definition of ξ we conclude that $u''(\xi_*,\alpha_*)\leq 0$. Let us first suppose that

$$u''(\xi_*,\alpha_*)=0. \tag{5.8}$$

Then, by the energy identity (1.7),

$$u(\xi_*,\alpha_*)=1.$$

We assert that (5.8) implies that

$$u'''(\xi_*, \alpha_*) > 0. \tag{5.9}$$

For if $u'''(\xi_*, \alpha_*) < 0$, then $u'' > 0$ and $u' < 0$ in a left-neighborhood of ξ_*, which contradicts the definition of ξ_*. If $u'''(\xi_*, \alpha_*) = 0$, it follows from uniqueness that $u(x) = 1$ for all $x \in \mathbf{R}$, which contradicts the condition at $x = 0$. Therefore, (5.9) holds (see also [21, Lemma 3.10]).

To complete the proof of Lemma 5.7, we need the following lemma in which we establish continuity of ξ at α_* under the above conditions.

Lemma 5.8. *Suppose that for some $\alpha_0 > 0$ we have*

$$u(\xi(\alpha_0), \alpha_0) = 1, \quad u''(\xi(\alpha_0), \alpha_0) = 0, \quad \textit{and} \quad u'''(\xi(\alpha_0), \alpha_0) > 0.$$

Then

$$\xi(\alpha) \to \xi(\alpha_0) \quad \textit{as } \alpha \to \alpha_0.$$

Accepting Lemma 5.8 for the moment, we conclude that

$$u'''(\xi(\alpha), \alpha) \to u'''(\xi(\alpha_*), \alpha_*) \text{ as } \alpha \to \alpha_*, \quad \alpha \in \mathcal{T}.$$

Because $u'''(\xi(\alpha), \alpha) < 0$ for all $\alpha \in \mathcal{T}$, it follows that

$$u'''(\xi(\alpha_*), \alpha_*) \leq 0, \tag{5.10}$$

which contradicts (5.9). Thus, (5.8) cannot be true and we conclude that

$$u''(\xi(\alpha_*), \alpha_*) < 0. \tag{5.11}$$

It follows from (5.11) and the energy identity (1.7) that

$$\textit{either} \quad u(\xi_*, \alpha_*) > 1 \quad \textit{or} \quad u(\xi_*, \alpha_*) < 1$$

and that $\xi(\alpha)$ is continuous at $\alpha = \alpha_*$. Hence, by continuous dependence on initial data, if $u(\xi_*, \alpha_*) < 1$, then $u(\xi(\alpha), \alpha) < 1$ for α in a right-neighborhood of α_*. Since $(\alpha_*, \infty) \subset \mathcal{T}$, the definition of $\mathcal{T}$ implies that this is impossible. Thus,

$$u(\xi_*, \alpha_*) > 1. \tag{5.12}$$

Finally, regarding u''', we must have equality in (5.10). For if

$$u'''(\xi(\alpha_*), \alpha_*) < 0,$$

then continuity implies that $\alpha_* > \inf \mathcal{T}$, a contradiction. Therefore,

$$u'''(\xi(\alpha_*), \alpha_*) = 0. \tag{5.13}$$

We conclude from (5.13) that the solution $u(x, \alpha_*)$ of problem (2.1) satisfies the properties (2.3) at $x = \xi(\alpha_*)$ and, thus, yields a periodic solution of which, by (5.12), the maximum is greater than 1. This completes the proof of Lemma 5.7.

The proof of Theorem 5.4 is complete once we have proved Lemma 5.8.

Proof of Lemma 5.8. Fix $\varepsilon > 0$ and small. Then by the assumptions on $u(\cdot, \alpha_0)$ there exists a $\delta > 0$ such that

$$u(\xi_0 - \varepsilon, \alpha_0) < 1 - 2\delta, \quad u(\xi_0 + \varepsilon, \alpha_0) > 1 + 2\delta, \tag{5.14a}$$

and

$$u'(x,\alpha_0) > \delta \quad \text{for all } x \in [0, \xi_0 - \varepsilon]. \tag{5.14b}$$

We wish to prove that there exists a $\nu > 0$ such that if $|\alpha - \alpha_0| < \nu$, then $u'(\cdot, \alpha)$ has a zero on $(\xi_0 - \varepsilon, \xi_0 + \varepsilon)$.

By the continuous dependence of solutions on initial data it follows from (5.14) that there exists a $\nu_1 > 0$ such that

$$u(\xi_0-\varepsilon,\alpha) < 1-\delta, \quad u(\xi_0+\varepsilon,\alpha) > 1+\delta, \text{ and } u'(x,\alpha) > 0 \text{ for all } x \in [0, \xi_0-\varepsilon] \tag{5.15}$$

if $|\alpha - \alpha_0| < \nu_1$ so that

$$\xi(\alpha) > \xi_0 - \varepsilon \quad \text{if } |\alpha - \alpha_0| < \nu_1. \tag{5.16}$$

To show that $\xi(\alpha) < \xi_0 + \varepsilon$ for α sufficiently close to α_0, it is sufficient to prove that

$$\tau(\alpha) \to \tau(\alpha_0) = 1 \quad \text{as } \alpha \to \alpha_0, \tag{5.17}$$

where we recall that $\tau(\alpha) = u(\xi(\alpha), \alpha)$. For (5.17) implies that there exists a $\nu_2 > 0$ such that

$$\tau(\alpha) < 1 + \delta \quad \text{if } |\alpha - \alpha_0| < \nu_2,$$

and, because $u' > 0$ on $(0, \xi)$, we conclude from (5.15) that

$$\xi(\alpha) < \xi_0 + \varepsilon \quad \text{if } |\alpha - \alpha_0| < \nu = \min\{\nu_1, \nu_2\}. \tag{5.18}$$

Thus, (5.15) and (5.18) yield the continuity of $\xi(\alpha)$ at α_0.

Let us now prove (5.17). Let $z_0(t) = z(t, \alpha_0)$ be the solution of problem (2.6) which corresponds to $u(x, \alpha_0)$. Then

$$z_0(t) \to 0 \quad \text{and} \quad \sqrt{z_0(t)}z_0''(t) \to A \quad \text{as } t \to 1^-,$$

where by assumption $A = 2u'''(\xi(\alpha_0), \alpha_0)$ is a positive constant. It is readily shown that this implies that the function $y_0(t) = z_0^{3/4}(t)$ has the properties

$$\frac{y_0(t)}{1-t} \to B \quad \text{and} \quad y_0'(t) \to -B \quad \text{as } t \to 1^-, \tag{5.19}$$

where $B = \frac{3}{2}\sqrt{A}$.

We can write the equation (2.6a) for z as

$$(z^{-1/4}z')' = \frac{z-f}{\gamma z^{5/4}},$$

so that, since $f(t) \geq 0$ for all $t \geq 0$, the function $y(t) = z^{3/4}(t)$ satisfies

$$y'' \leq \frac{3}{4\gamma} y^{-1/3}. \tag{5.20}$$

Fix $\rho \in (0, 1)$. Then $y_0(1 - \rho) > 0$ and it follows from the continuous dependence on initial data on $[0, \rho]$ that there exists a $\vartheta_1 > 0$ such that $\tau(\alpha) > 1 - \rho$ when $|\alpha - \alpha_0| < \vartheta_1$. Since ρ may be chosen as small as we wish, we conclude that

$$\liminf_{\alpha \to a_0} \tau(\alpha) \geq 1. \tag{5.21a}$$

It remains to prove that

$$\limsup_{\alpha\to a_0} \tau(\alpha) \le 1. \tag{5.21b}$$

Fix $\varepsilon > 0$ and $t_0 \in (0,1)$. By (5.19), it is possible to choose t_0 so close to 1 that

$$y_0'(t_0) \le -\frac{\sqrt{3}}{2}B \quad \text{and} \quad y_0(t_0) \le \frac{B}{8}\,\varepsilon.$$

By continuity we can find a constant $\vartheta_2 > 0$ so small that if $|\alpha - \alpha_0| < \vartheta_2$, then

$$y'(t_0) \le -\frac{B}{2} \quad \text{and} \quad 0 < y(t_0) \le \frac{B}{4}\varepsilon. \tag{5.22}$$

Thus, in a neighborhood of t_0 we have $y' < 0$, so that when we multiply (5.20) by y' we obtain

$$(y'^2)' \ge \frac{9}{4\gamma}(y^{2/3})'$$

for $t > t_0$ as long as $y > 0$ and $y' < 0$. This yields, upon integration over (t_0, t),

$$\begin{aligned} y'^2(t) &\ge y'^2(t_0) + \frac{9}{4\gamma}\{y^{2/3}(t) - y^{2/3}(t_0)\} \\ &\ge y'^2(t_0) - \frac{9}{4\gamma}y^{2/3}(t_0) \\ &\ge \frac{B^2}{4} - \frac{9}{4\gamma}\Big(\frac{B\varepsilon}{4}\Big)^{2/3} > \frac{B^2}{16}, \end{aligned}$$

if we choose $\varepsilon < \varepsilon_0 = \frac{1}{2}(\frac{\gamma}{3})^{3/2}B^2$. Therefore,

$$y'(t) < -\frac{B}{4} \quad \text{for } t_0 \le t \le \tau.$$

Thus, when $0 < \varepsilon < \varepsilon_0$, integration over (t_0, τ) yields

$$\tau \le t_0 + \frac{4}{B}\,y(t_0) < t_0 + \varepsilon < 1 + \varepsilon,$$

where we have used (5.22). Since ε can be chosen arbitrarily small, this proves (5.21b) and the proof of Lemma 5.8 is complete.

We conclude this section with a few observations about the existence of periodic solutions with amplitude greater than 1 when $0 < \mu < 1$.

Because the initial data, and hence the solution u of problem 2.1, depend continuously on μ, it is evident from the proof of Theorem 5.1 that there exist periodic solutions with amplitude greater than 1 when μ is sufficiently small. In the following theorem we shall show that this is no longer true when $\mu \ge \frac{4}{9}$.

THEOREM 5.9. *If $\mu \ge \frac{4}{9}$ and $\gamma > 0$ is arbitrary, then there exists no periodic solution u for which*

$$\max\{|u(x)| : x \in \mathbf{R}\} > 1.$$

Proof. Let $\mu \in (0,1)$ and suppose that z corresponds to a periodic solution such that

$$\tau = \sup\{u(x) : x \in \mathbf{R}\} > 1.$$

Then (1.7) implies that

$$\tau > \sqrt{1+\sqrt{\mu}}.$$

From Lemma 4.1(b) we know that

$$z'(t) < 0 \quad \text{and} \quad 0 < t < \tau,$$

and by multiplying equation (2.6a) by $\frac{3}{4}z^{-5/4}$ we obtain

$$(z^{3/4})'' = \frac{3}{4\gamma}\frac{z - f_\mu}{z^{5/4}} \quad \text{for } 0 < t < \tau, \tag{2.10}$$

where

$$f_\mu = \frac{1}{2}\{(1-t^2)^2 - \mu\}.$$

Let us denote the zeros of f_μ by a and b:

$$a = \sqrt{1-\sqrt{\mu}} \quad \text{and} \quad b = \sqrt{1+\sqrt{\mu}}.$$

Because $\tau > b$, we can integrate (2.10) over $(0,b)$ to obtain

$$\frac{3}{4\gamma}\Big(\int_0^a + \int_a^b\Big)\frac{z-f_\mu}{z^{5/4}}\,dt = (z^{3/4})'\Big|_0^b < 0,$$

where the inequality is clear when we remember that $z'(0) = 0$ and $z'(b) < 0$. Thus, writing

$$I_1 = \int_0^a \frac{z-f_\mu}{z^{5/4}}\,dt \quad \text{and} \quad I_2 = \int_a^b \frac{z-f_\mu}{z^{5/4}}\,dt,$$

we have

$$I_1 + I_2 < 0. \tag{5.23}$$

Recall that $z > 0$ and $z' < 0$ on $(0,\tau)$. Hence

$$I_1 > -\int_0^a \frac{f_\mu}{z^{5/4}}\,dt > -\frac{1}{z^{5/4}(a)}\int_0^a f_\mu(t)\,dt \tag{5.24}$$

and

$$I_2 = \int_a^b \frac{z+|f_\mu|}{z^{5/4}}\,dt > \frac{1}{z^{5/4}(a)}\int_a^b |f_\mu(t)|\,dt > 0, \tag{5.25}$$

because $f_\mu < 0$ on (a,b). Putting (5.24) and (5.25) into (5.23) we find that

$$-\int_0^a f_\mu(t)\,dt + \int_a^b |f_\mu(t)|\,dt < 0,$$

or, equivalently,

$$\int_0^b f_\mu(t)\,dt > 0. \tag{5.26}$$

An elementary computation shows that (5.26) holds if and only if

$$\mu < \frac{4}{9}.$$

Thus, if $\mu \geq \frac{4}{9}$, there can be no periodic solution with maxima above 1.

6. Qualitative properties. In this section we prove several qualitative properties of periodic solutions. We begin with a convexity lemma and then we establish universal global bounds for periodic solutions. This is followed by an analysis of the behavior of periodic solutions as $\gamma \to 0$ (when $0 < \mu < 1$), as $\gamma \to \frac{1}{8}$ (when $\mu = 0$), and as $\gamma \to \infty$.

We begin with a convexity property.

LEMMA 6.1. *Let $u(x)$ be a periodic solution which has a single critical point between zeros and has the symmetry properties* (1.11). *Then*

$$u''(x) < 0 \quad \textit{when } u(x) > 0.$$

Proof. By Lemma 4.1(b), if $z(t)$ is the solution of problem (2.6) which corresponds to $u(x)$, then $z'(t) < 0$ for $0 < t < \tau$, and hence,

$$u''(x) = \frac{1}{2} z'(t(x)) < 0 \quad \text{for } 0 < x < \xi.$$

Since $u''(\xi) < 0$ by the energy identity, the assertion follows.

A remarkable feature of all single-bump periodic solutions is that they are bounded by a constant which does not depend on either γ or μ. This is shown in the next lemma.

LEMMA 6.2. *Let $0 \leq \mu < 1$ and $\gamma > 0$, and let $u(x)$ be a periodic solution that has the symmetry properties* (1.11). *Then*

$$|u(x)| < \sqrt{2} \quad \textit{for } x \in (-\infty, \infty).$$

Proof. Suppose that for some $a \in \mathbf{R}$ we have $|u(a)| \geq \sqrt{2}$. Without loss of generality we may assume that u has a maximum at $x = a$. Thus, we have

$$u(a) \geq \sqrt{2}, \quad u'(a) = 0, \quad u''(a) \leq -\sqrt{\frac{1-\mu}{2\gamma}}, \quad u'''(a) = 0, \tag{6.1}$$

where the upper bound for u'' follows from the energy identity (1.7). From (1.6) we see that $u^{iv}(a) < 0$.

Thus, as x increases above a, u''' and u'' decrease. Thus, $u''' < 0$ and $u'' < 0$ as long as $u^{iv} < 0$. Furthermore,

$$u^{iv} = u'' + u - u^3 < 0 \quad \text{as long as } u > 1.$$

Thus, if

$$b = \sup\{x > a : u > 1 \ \text{ on } \ [a, x)\},$$

then

$$u(b)=1,\quad u'(b)<0,\quad u''(b)<-\sqrt{\frac{1-\mu}{2\gamma}},\quad u'''(b)<0. \tag{6.2}$$

We claim that

$$u''(x)<-\sqrt{\frac{1-\mu}{2\gamma}}\quad\text{as long as } |u|<1. \tag{6.3}$$

Suppose that (6.3) does not hold. Then u''' must have a zero at a point where $|u|<1$. Let $y>b$ be the first zero of u'''. Then by (1.7),

$$-\gamma(u'')^2-(u')^2+\frac{1}{2}(1-u^2)^2=\frac{\mu}{2}\quad\text{at } y.$$

Because of our assumption that $|u(y)|<1$,

$$\{u''(y)\}^2<\frac{1-\mu}{2\gamma}.$$

This means that

$$u''(y)>-\sqrt{\frac{1-\mu}{2\gamma}}.$$

However, since $u'''<0$ on (b,y), it follows from (6.2) that

$$u''(y)<-\sqrt{\frac{1-\mu}{2\gamma}},$$

so we have a contradiction.

Thus, (6.3) holds. In particular, we have that $u''<0$ at the first zero of u and, by (1.11), at all zeros of u. This contradicts the fact that because u is odd, $u''=0$ whenever $u=0$. Thus, we must conclude that the assertion holds.

We now turn to a discussion of the behavior of periodic solutions for values of γ close to $\gamma=0$, or $\gamma=\frac{1}{8}$ when $\mu=0$, and for large values of γ.

LEMMA 6.3. *Let $\{\gamma_i\}$ be a sequence such that*

$$\gamma_i\searrow\theta=\begin{cases}\dfrac{1}{8} & \textit{when } \mu=0\\ 0 & \textit{when } 0<\mu<1\end{cases}\qquad \textit{as } i\to\infty.$$

For each $i\geq 1$, let u_i be a periodic solution corresponding to γ_i. Then

$$u_i(x)\to U(x)\quad \textit{as } i\to\infty,\ \textit{uniformly on compact intervals},$$

where

(i) *if $\mu=0$, then U is the unique monotone symmetric kink corresponding to $\gamma=\frac{1}{8}$;*

(ii) *if $0<\mu<1$, then U is the unique periodic solution of the FK equation with energy μ.*

Proof. Let $\alpha_i=u_i'(0)$. If $0<\mu<1$, then Lemmas 3.3 and 3.4 imply that

$$\sqrt{\frac{1-\mu}{4}}\leq\alpha_i\leq\sqrt{\frac{1-\mu}{2}} \tag{6.4}$$

for i sufficiently large and $\gamma_i - \theta > 0$ sufficiently small. If $\mu = 0$, then (6.4) follows from Lemmas 5.2, 5.5, and 5.6. Hence, there exists a convergent subsequence, which we also denote by $\{\gamma_i\}$, such that

$$\alpha_i \to \hat{\alpha} \quad \text{as } i \to \infty,$$

where $\hat{\alpha}$ satisfies (6.4).

We consider the cases $\mu = 0$ and $0 < \mu < 1$ in succession.

Case I. Let $\mu = 0$ and let $\alpha_0 = U'(0)$, where U is the kink for $\gamma = \theta = \frac{1}{8}$. Suppose that $\hat{\alpha} > \alpha_0$. Then, by [21, Lemma 3.6 and Theorem 3.7],

$$u\left(\xi\left(\hat{\alpha}, \frac{1}{8}\right), \hat{\alpha}, \frac{1}{8}\right) > 1 \quad \text{and} \quad u''\left(\xi\left(\hat{\alpha}, \frac{1}{8}\right), \hat{\alpha}, \frac{1}{8}\right) < 0. \tag{6.5}$$

Since by Theorem 4.5, $u(\cdot, \hat{\alpha}, \frac{1}{8})$ cannot be a periodic solution, it also follows that

$$u'''\left(\xi\left(\hat{\alpha}, \frac{1}{8}\right), \hat{\alpha}, \frac{1}{8}\right) \neq 0. \tag{6.6}$$

The inequality in (6.5) implies that the function $\xi(\alpha, \gamma)$ is continuous at $(\hat{\alpha}, \frac{1}{8})$ so, by the continuous dependence of $u(\cdot, \alpha, \gamma)$ on α and γ, it follows that for i large enough,

$$u'''(\xi(\alpha_i, \gamma_i), \alpha_i, \gamma_i) \neq 0 \tag{6.7}$$

as well. However, since $u(\cdot, \alpha_i, \gamma_i)$ is a periodic solution for every i, we must have

$$u'''(\xi(\alpha_i, \gamma_i), \alpha_i, \gamma_i) = 0$$

for every i, which contradicts (6.7).

If $\hat{\alpha} < \alpha_0$, then by [21, Lemma 3.6 and Theorem 3.7],

$$u\left(\xi\left(\hat{\alpha}, \frac{1}{8}\right), \hat{\alpha}, \frac{1}{8}\right) < 1 \quad \text{and} \quad u''\left(\xi\left(\hat{\alpha}, \frac{1}{8}\right), \hat{\alpha}, \frac{1}{8}\right) < 0.$$

It follows from Theorem 4.4 that (6.6) holds again and, as before, we arrive at a contradiction.

Thus, for the subsequence we have $\alpha_i \to \alpha_0$ as $i \to \infty$. Because the limit is uniquely determined, it follows that the entire sequence $\{\alpha_i\}$ converges to α_0. Therefore, $u_i \to U$ uniformly on compact sets of the form $[0, L]$, $L > 0$.

Case II. Let $0 < \mu < 1$ and let $\alpha_\mu = U'(0)$, where U is the periodic solution of the FK equation with energy equal to μ. From the energy identity (1.7) in which we set $\gamma = 0$, we conclude that

$$\alpha_\mu = \sqrt{\frac{1-\mu}{2}}.$$

It follows from Corollary 3.6 that

$$\hat{\alpha} \leq \alpha_\mu.$$

In the remainder of the proof we shall show that

$$\hat{\alpha} \geq \alpha_\mu$$

as well. This proves that $\alpha_i \to \alpha_\mu$ as $i \to \infty$, and therefore $u_i \to U$ uniformly on compact sets, as asserted.

We shall show that for each $\varepsilon > 0$ there exists a $\gamma_\varepsilon > 0$ such that if $0 < \gamma < \gamma_\varepsilon$ and $u(\cdot, \alpha, \gamma)$ is a periodic solution, then $\alpha > \alpha_\mu - \varepsilon$ so

$$\liminf_{i \to \infty} \alpha_i \geq \alpha_\mu.$$

Remember that the initial conditions for u are

$$u(0) = 0, \quad u'(0) = \alpha, \quad u''(0) = 0, \quad u'''(0) = \beta(\alpha) = \frac{\alpha^2 - \alpha_\mu^2}{2\alpha\gamma}.$$

Because $\beta'(\alpha) > 0$ it follows that

$$\beta(\alpha) \leq \beta(\alpha_\mu - \varepsilon) = -\frac{\delta(\varepsilon)}{\gamma} \quad \text{for } 0 < \alpha \leq \alpha_\mu - \varepsilon,$$

where $\delta(\varepsilon) \sim \varepsilon$ as $\varepsilon \to 0$.

We now proceed as in the proof of Lemma 3.3. Because $u(0) = 0$ and $u'''(0) < 0$, it follows that $u < 1$ and $u''' < 0$ in a neighborhood of the origin. As long as these inequalities do not change it follows from the equation for u that

$$u^{iv}(x) < \frac{\alpha}{\gamma} x, \tag{6.8a}$$

$$u'''(x) < -\frac{\delta}{\gamma} + \frac{\alpha}{2\gamma} x^2, \tag{6.8b}$$

$$u''(x) < -\frac{\delta}{\gamma} x + \frac{\alpha}{6\gamma} x^3, \tag{6.8c}$$

$$u'(x) < \alpha - \frac{\delta}{2\gamma} x^2 + \frac{\alpha}{24\gamma} x^4. \tag{6.8d}$$

Set

$$x_\gamma = \gamma^{1/4}.$$

Then, when $\gamma < \gamma_1 = (2\delta/\alpha_\mu)^2$, it follows from (6.6b) that $u''' < 0$ on $(0, x_\gamma]$. Moreover, the right-hand side of (6.6d) will be negative at x_γ if $\gamma < \gamma_2 = \{12\delta/(25\alpha_\mu)\}^2$.

Thus, if we set $\gamma_\varepsilon = \min\{\gamma_1, \gamma_2\}$ and denote as usual the first zero of u' by ξ, then

$$\xi \in (0, x_{\gamma_\varepsilon}) \quad \text{and} \quad u'''(\xi) < 0 \quad \text{if } 0 < \gamma < \gamma_\varepsilon.$$

This means that if $\alpha \leq \alpha_\mu - \varepsilon$ and $\gamma \in (0, \gamma_\varepsilon)$, then $u(\cdot, \alpha, \gamma)$ cannot be a periodic solution. Therefore, if it is given that $u(\cdot, \alpha, \gamma)$ is a periodic solution, then we must conclude that $\alpha > \alpha_\mu - \varepsilon$, and the proof is complete.

To determine the behavior of periodic solutions as $\gamma \to \infty$, we first need an upper and a lower bound for the slope at the origin.

Lemma 6.4. *Let $0 \leq \mu < 1$ and $\gamma > 0$, and let $u(x)$ be a periodic solution which satisfies* (1.11). *Then*

$$u'(0) \leq \{8(1-\mu)\log 2\}^{1/4} \gamma^{-1/4} \quad \textit{for } \gamma > \theta,$$

where $\theta = 0$ if $0 < \mu < 1$ and $\theta = \frac{1}{8}$ if $\mu = 0$.

Proof. It is convenient to prove this lemma using the function $z(t)$ introduced in section 3. We then need to show that

$$z(0) \leq \{8(1-\mu)\log 2\}^{1/2}\gamma^{-1/2}. \tag{6.9}$$

It follows from (2.6a) that

$$zz'' = \frac{(z')^2}{4} + \frac{1}{\gamma}\{z - f_\mu(t)\} > -\frac{1-\mu}{2\gamma} \quad \text{for } 0 \leq t < \sqrt{2}.$$

Because $\tau < \sqrt{2}$ by Lemma 6.2, it follows that

$$z'' > -\frac{1-\mu}{2\gamma z} \quad \text{for } 0 \leq t < \tau.$$

We multiply by $z' < 0$, integrate over $(0, t)$, and obtain

$$z'(t) > -\left\{\frac{1-\mu}{\gamma}\log\frac{z(0)}{z(t)}\right\}^{1/2} \quad \text{for } 0 \leq t < \tau. \tag{6.10}$$

Let

$$t_0 = \sup\left\{t > 0 : z > \frac{1}{2}z(0) \;\text{ on }\; [0,t)\right\}.$$

Then

$$0 < t_0 < \sqrt{2} \quad \text{and} \quad \frac{z(0)}{z(t)} \leq 2 \quad \text{for } 0 \leq t \leq t_0.$$

Hence, by (6.10),

$$z'(t) > -\left(\frac{(1-\mu)\log 2}{\gamma}\right)^{1/2} \quad \text{for } 0 < t < t_0$$

and we obtain, after an integration over $(0, t_0)$,

$$z(t_0) - z(0) > -\left(\frac{(1-\mu)\log 2}{\gamma}\right)^{1/2} t_0. \tag{6.11}$$

Since $t_0 < \sqrt{2}$, this implies that

$$\frac{1}{2}z(0) < \left(\frac{(1-\mu)\log 2}{\gamma}\right)^{1/2},$$

from which (6.9) follows.

Next we establish a lower bound for $u'(0)$.

LEMMA 6.5. *We have*

$$u'(0) > \frac{1}{5}\sqrt{1-\mu}\,\gamma^{-1/4} \quad \textit{for } \gamma > \left(\frac{2}{5}\right)^4.$$

Proof. In light of the upper bound obtained in Lemma 6.3, we scale the variables and set

$$s = \frac{x}{\gamma^{1/4}} \quad \text{and} \quad v(s) = u(x). \tag{6.12}$$

We then obtain the problem

$$\begin{cases} v^{iv} = \dfrac{v''}{\sqrt{\gamma}} + v - v^3, & \text{(6.13a)} \\ v(0) = 0, \quad v''(0) = 0, & \text{(6.13b)} \\ v'(0) = \omega, \quad v'''(0) = \dfrac{1}{2\omega}\Big(\dfrac{\omega^2}{\sqrt{\gamma}} - \dfrac{1-\mu}{2}\Big) & \text{(6.13c)} \end{cases}$$

in which

$$\omega = \alpha\gamma^{1/4}$$

and we need to prove that

$$\omega > \frac{1}{5}\sqrt{1-\mu}.$$

Suppose, to the contrary, that

$$\omega \leq \frac{1}{5}\sqrt{1-\mu}.$$

Then for $\gamma > (2/5)^4$, we have

$$v'''(0) < -\frac{1-\mu}{8\omega}.$$

As long as $v > 0$ and $v''' < 0$, we have

$$v(s) < \omega s$$

and

$$v^{iv}(s) < \omega s,$$

which yields, upon integration over $(0, s)$,

$$v'''(s) < -\frac{1-\mu}{8\omega} + \frac{1}{2}\omega s^2. \tag{6.14}$$

One verifies that the right-hand side of (6.14) is negative for all $s \in [0, 1]$. Two more integrations yield

$$v'(s) < \omega - \frac{1-\mu}{16\omega}s^2 + \frac{1}{24}\omega s^4.$$

It follows that the first zero $\sigma = \sigma(\omega, \gamma)$ of v' has the properties

$$0 < \sigma < 1 \quad \text{and} \quad v'''(\sigma) < 0,$$

so v, and hence u, cannot be periodic solutions, a contradiction.

With the lower bound we now have in hand we can return to the argument used in the proof of Lemma 6.1 to obtain a lower bound for the maximum of $|u(x)|$ on $\mathbf{R}$.

LEMMA 6.6. *Let $u(x,\gamma)$ be a periodic solution. Then*

$$\max\{|u(x,\gamma)| : x \in \mathbf{R}\} > \frac{1}{50}\sqrt{\frac{1-\mu}{\log 2}} \quad \textit{if } \gamma > \left(\frac{2}{5}\right)^4.$$

Proof. If u is a periodic solution, then by Lemma 6.5,

$$z(0) > \frac{1-\mu}{25\sqrt{\gamma}} \quad \text{if } \gamma > \left(\frac{2}{5}\right)^4.$$

Therefore, by (6.11),

$$\left(\frac{(1-\mu)\log 2}{\gamma}\right)^{1/2} t_0 > \frac{1-\mu}{50\sqrt{\gamma}}.$$

This means that

$$t_0 > \frac{1}{50}\sqrt{\frac{1-\mu}{\log 2}}.$$

Now, because

$$\max\{|u(x,\gamma)| : x \in \mathbf{R}\} = \tau(\gamma) > t_0,$$

the assertion follows.

From Lemmas 6.4 and 6.5 we conclude that if u is a periodic solution which satisfies (1.11), then for γ large enough,

$$\frac{1}{8}\sqrt{1-\mu} < \omega < \{8(1-\mu)\log 2\}^{1/4}.$$

Let $\{\gamma_i\}$ be a sequence tending to infinity and let u_i be a corresponding sequence of periodic solutions, with initial slopes α_i. Let v_i and ω_i be the solutions of problem (6.13) corresponding to these periodic solutions. Then by compactness there exists a subsequence, which we also denote by $\{\omega_i\}$, which converges to a number $\omega^* < \infty$ as $i \to \infty$. Plainly, it must be the case that $v_i \to V$, uniformly on compact sets, as $i \to \infty$, where V satisfies

$$\begin{cases} V^{iv} = V - V^3, & \text{(6.15a)} \\ V(0) = 0, \quad V''(0) = 0, & \text{(6.15b)} \\ V'(0) = \omega^*, \quad V'''(0) = -\dfrac{1-\mu}{4\omega^*}. & \text{(6.15c)} \end{cases}$$

We assert that the sequence of maxima $\{\sigma_i\}$, where $\sigma_i = \sigma(\omega_i, \gamma_i)$, remains bounded:

$$\limsup_{i\to\infty} \sigma(\omega_i, \gamma_i) < \infty.$$

For if not, then there exists a subsequence along which σ_i tends to infinity. Hence,

$$V'(s) > 0 \quad \text{for } 0 \le s < \infty.$$

An argument like the one used in the proof of Lemma 2.1 shows that this is impossible. Therefore, the sequence $\{\sigma_i\}$ is bounded and there exists a subsequence which converges to some $\sigma^* < \infty$ as $i \to \infty$. Since $v_i'''(\sigma_i) = 0$ for every i, it easily follows that $V'''(\sigma^*) = 0$, so that V is a periodic solution of problem (6.15).

Thus we have the following lemma.

LEMMA 6.7. *Let* $0 \le \mu < 1$. *Suppose that* $\{\gamma_i\}$ *is a sequence which tends to infinity and* $\{u_i\}$ *is a sequence of periodic solutions which correspond to* γ_i *and which each satisfy* (1.11). *Then there exists a subsequence and a periodic solution* V *of problem* (6.15) *which also satisfies* (1.11) *such that*

$$u_i(\gamma_i^{1/4}s, \gamma_i) \to V(s) \quad \text{as } i \to \infty, \tag{6.16}$$

uniformly on compact sets.

As a by-product, the above argument yields the existence of periodic solutions V of problem (6.15) which satisfy (1.11) for every $\mu \in [0, 1)$. This result can also be proved by means of the method used in sections 3 and 5. Since the proofs are very close to those already presented, we omit them here. Summarizing, we have the following result.

THEOREM 6.8. *If* $0 \le \mu < 1$, *then problem* (6.15) *has a periodic solution* V_1 *such that*

$$\max\{|V_1(x)| : x \in \mathbf{R}\} < 1.$$

If $\mu = 0$, *then problem* (6.15) *has a periodic solution* V_2 *which satisfies* (1.11) *such that*

$$\max\{|V_2(x)| : x \in \mathbf{R}\} \in (1, \sqrt{2}).$$

Acknowledgment. It is a pleasure to thank J. Serrin for numerous stimulating discussions. We thank the School of Mathematics, as well as the Institute of Mathematics and its Application, at the University of Minnesota and the University of Pittsburgh for providing hospitality during the preparation of this paper. Finally the authors are much indebted to the referee, whose questions and comments led to a substantially improved manuscript.

REFERENCES

[1] N. N. Akhmediev, A. V. Buryak, and M. Karlsson, *Radiationless optical solitons with oscillating tails*, Optics Comm., 110 (1994), pp. 540–544.

[2] C. J. Amick and J. F. Toland, *Homoclinic orbits in the dynamic phase space analogy of an elastic strutt*, European J. Appl. Math., 3 (1992), pp. 97–114.

[3] D. G. Aronson and H. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[4] B. Buffoni, A. R. Champneys, and J. F. Toland, *Bifurcation and coalescence of a plethora of homoclinic orbits for a hamiltonian system*, J. Dynam. Differential Equations, 8 (1996), pp. 221–281.

[5] E. Bodenschatz, M. Kaiser, L. Kramer, W. Pesch, A. Weber, and W. Zimmerman, *Patterns and Defects in Liquid Crystals in New Trends in Nonlinear Dynamics and Pattern Forming Phenomena: The Geometry of Nonequilibrium*, P. Coullet and P. Huerre, eds., NATO ASI Series, Plenum Press, New York, 1990, p. 111.

[6] B. Buffoni and J. F. Toland, *Global existence of homoclinic periodic orbits for a class of autonomous Hamiltonian systems*, J. Differential Equations, 118 (1995), pp. 104–120.

[7] P. Collet and J.-P. Eckmann, *Instabilities and Fronts in Extended Systems*, Princeton University Press, Princeton, NJ, 1980.

[8] P. Coullet, C. Elphick, and D. Repaux, *Nature of spatial chaos*, Phys. Rev. Lett., 58 (1987), pp. 431–434.

[9] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Mod. Phys., 65 (1993), pp. 851–1112.

[10] B. D. Coleman, M. Marcus, and V. J. Mizel, *On the thermodynamics of periodic phases*, Arch. Rational Mech. Anal., 117 (1992), pp. 321–347.

[11] G. T. Dee and W. van Saarloos, *Bistable systems with propagating fronts leading to pattern formation*, Phys. Rev. Lett., 60 (1988), pp. 2641–2644.

[12] R. L. Devaney, *Homoclinic orbits in Hamiltonian systems*, J. Differential Equations, 21 (1976), pp. 431–438.

[13] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling wave solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–362.

[14] R. M. Hornreich, M. Luban, and S. Shtrikman, *Critical behaviour at the onset of* **k***-space instability on the* λ *line*, Phys. Rev. Lett., 35 (1975), p. 1678.

[15] W. D. Kalies, W. Kwapisz, and R. C. A. M. van der Vorst, *Homotopy classes for stable connections between Hamiltonian saddle-focus equilibria*, Comm. Math. Phys., to appear.

[16] A. Kolmogorov, I. Petrovski, and N. Piscounov, *Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bull. Univ. Moskou, Ser. Internat., Sec. A, 1 (1937), pp. 1–25.

[17] A. Leizarowitz and V. J. Mizel, *One dimensional infinite-horizon variational problems arising in continuum mechanics*, Arch. Rational Mech. Anal., 106 (1989), pp. 161–194.

[18] M. Marcus, *Uniform estimates for a variational problem with small parameters*, Arch. Rational Mech. Anal., 117 (1992), pp. 321–347.

[19] P. J. McKenna and W. Walter, *Travelling waves in a suspension bridge*, SIAM J. Appl. Math., 50 (1990), pp. 703–715.

[20] J. A. Powell, A. C. Newell, and C. K. R. T. Jones, *Competition between generic and nongeneric fronts in envelope equations*, Phys. Rev. A.(3), 44 (1991), pp. 3636–3652.

[21] L. A. Peletier and W. C. Troy, *Spatial patterns described by the extended Fisher–Kolmogorov (EFK) equation: Kinks*, Differential Integral Equations, 8 (1995), pp. 1279–1304.

[22] L. A. Peletier and W.C. Troy, *A topological shooting method and the existence of kinks of the extended Fisher–Kolmogorov equation*, Topol. Methods Nonlinear Anal., 6 (1996), pp. 331–355.

[23] L. A. Peletier and W.C. Troy, *Chaotic spatial patterns described by the extended Fisher–Kolmogorov equation*, J. Differential Equations, 129 (1996), pp. 458–508.

[24] L. A, Peletier, W. C. Troy, and R. C. A. M. van der Vorst, *Stationary solutions of a fourth order nonlinear diffusion equation*, Differentsial'nye Uravneniya, 31 (1995), pp. 327–337 (in Russian) and Differential Equations, 31 (1995), pp. 301–314 (in English).

[25] W. van Saarloos, *Front propagation into unstable states: Marginal stability as a dynamical mechanism for velocity selection*, Phys. Rev. A.(3), 37 (1988), pp. 211–229.

[26] W. van Saarloos, *Front propagation into unstable states.* II. *Linear versus nonlinear marginal stability and rate of convergence*, Phys. Rev. A.(3), 39 (1989), pp. 6367–6390.

[27] W. Zimmerman, *Propagating fronts near a Lifschitz point*, Phys. Rev. Lett., 66 (1991), p. 1546.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1354–1370, November 1997

005

STABILITY ESTIMATES IN AN INVERSE PROBLEM FOR A THREE-DIMENSIONAL HEAT EQUATION*

SERGIO VESSELLA†

Abstract. The problem of determining an insulating body D contained in a conducting one Ω is studied. If at an initial time the temperature is zero and increasing temperature is assigned on the boundary of Ω then the knowledge of the flux on a portion of $\partial\Omega$ for a finite interval of time determines D. A logarithmic stability estimate is found if some a priori assumptions are given on D.

Key words. stability, inverse problem

AMS subject classifications. 35K05, 35R30

PII. S0036141095294262

1. Introduction.

Consider a three-dimensional heat conducting body Ω, and suppose that it contains an insulating three-dimensional body D. We want to determine D by assigning the initial temperature of Ω, the temperature on the boundary for an interval of time $[0, t_1]$, and measuring the flux of temperature on a portion Σ of $\partial\Omega$ in a subinterval of $[0, t_1]$.

Denote by $u(x, t)$ the temperature in a point x at time t and suppose that the temperature of Ω is initially zero (or, more generally, that $\Delta u(x, 0) \geq 0$). If φ is the temperature on $\partial\Omega$ then u solves the following boundary value problem:

$$\frac{\partial u}{\partial t} - \Delta u = 0 \text{ in } \left(\Omega \setminus \overline{D}\right) \times [0, t_1]\,, \tag{1}$$

$$u = \varphi \text{ on } \partial\Omega \times [0, t_1]\,, \tag{2}$$

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial D \times [0, t_1]\,, \tag{3}$$

$$u\,(x, 0) = 0 \text{ for } x \in \Omega \setminus D, \tag{4}$$

where $\mathbf{n}$ denotes the unit outer normal to $\partial\,(\Omega \setminus D)$.

The problem we study consists of assigning an opportune input φ on $\partial\Omega \times [0, t_1]$ in such a way that the flux $\frac{\partial u}{\partial \mathbf{n}}$ on $\Sigma \times [t_0, t_1]$, where Σ is a nonempty open set of $\partial\Omega$, $t_0 \in (0, t_1)$, determines D.

It is simple to prove that uniqueness holds for a large class of domains D if the input φ satisfies the monotonicity condition

$$\frac{\partial\varphi\,(x, t)}{\partial t} > 0 \text{ for } (x, t) \in \partial\Omega \times [0, t_1]\,. \tag{5}$$

In fact, let $\mathcal{P}$ be the class of domains D that are C^2 and such that $\overline{D} \subset$int(Ω) and $\Omega \setminus \overline{D}$ is connected. Suppose that there exist two domains $D_1, D_2 \in \mathcal{P}$ such that

[*] Received by the editors November 30, 1995; accepted for publication (in revised form) June 25, 1996.

http://www.siam.org/journals/sima/28-6/29426.html

[†] Dipartimento di Matematica per le Decisioni Economiche, Finanziarie, Attuariali e Sociali (DIMADEFAS), Universitá degli Studi di Firenze, Via C. Lombroso 6/17, 50134 Firenze, Italy.

$\frac{\partial u_1}{\partial \mathbf{n}} = \frac{\partial u_2}{\partial \mathbf{n}}$ on $\Sigma \times [t_0, t_1]$. Let I be the connected component of $\overline{\Omega} \setminus (D_1 \cup D_2)$ that contains $\partial\Omega$; uniqueness for the parabolic Cauchy problem and analytic continuation in the space variables give $u_1 \equiv u_2$ in $I \times [t_0, t_1]$. Therefore, if, for instance, $D_2 \setminus D_1 \neq \emptyset$, then, by the inclusion $\partial\left((\Omega \setminus I) \setminus D_1\right) \subset (\partial I \cap \partial D_2) \cup \partial D_1$ and by the divergence theorem, we have for every $t \in (t_0, t_1)$

$$\int\limits_{(\Omega \setminus I) \setminus D_1} \frac{\partial u_1\,(x,t)}{\partial t} dx = \int\limits_{(\partial I) \cap (\partial D_2)} \frac{\partial u_1\,(x,t)}{\partial \mathbf{n}} ds = \int\limits_{(\partial I) \cap (\partial D_2)} \frac{\partial u_2\,(x,t)}{\partial \mathbf{n}} ds = 0,$$

but the maximum principle applied to $\frac{\partial u_1}{\partial t}$ yields that the first integral is strictly positive, which leads to a contradiction.

We are mainly interested in studying the stability, that is, the continuous dependence of D on the measured flux; to this purpose we assume that ∂D is connected and is locally a graphic of $\mathrm{C}^{2,\alpha}$ functions whose norms are bounded and $\operatorname{dist}(D, \partial\Omega) \geq L_0 > 0$. If this a priori information is available, (5) is satisfied and the following error bound on the measured flux of temperature holds:

$$\sup_{\Sigma \times [t_0, t_1]} \left| \frac{\partial u_1}{\partial \mathbf{n}} - \frac{\partial u_2}{\partial \mathbf{n}} \right| \leq \varepsilon.$$

Then (see Theorem 2.4)

$$\delta_{\mathcal{H}}\,(D_1, D_2) = O\left(\left| \ln \varepsilon \right|^{-s} \right),$$

where $\mathrm{D}_i, i = 1, 2$ is the domain corresponding to u_i, $\delta_{\mathcal{H}}\,(D_1, D_2)$ is the Hausdorff distance between D_1 and D_2, and s is a positive number.

We want to point out that our results can be extended simply to the case where we assign the flux of temperature, ϕ, on the boundary of Ω in the interval of time $[0, t_1]$ and we measure the temperature on Σ in a subinterval of $[0, t_1]$. If $\frac{\partial \phi}{\partial t} > 0$ and the initial temperature is zero then uniqueness and logarithmic stability remain valid.

An interesting open problem is the one where the body D is not perfectly insulating. In this case the problem that could be studied is the determination of the coefficient a in the equation $u_t - \operatorname{div}(a \nabla u) = 0$ in $\Omega \times [0, T]$, where $a = a_0 + b\chi_D$ (χ_D is the indicator function of D), a_0 and b are known positive constants, and D is the unknown to be determined, again, by a pair of Cauchy data. In [6] the authors consider the problem of determining the nonconstant function b and the domain D (in the case where a_0 is also a nonconstant function) from the whole Dirichlet-to-Neumann map.

In our opinion there is some similarity in the mathematical formulation between our problem and the problem of cracks determination. In these problems one wants to determine an unknown electrically insulating surface (or a curve in the two-dimensional version) which lies in the electrically conducting body from electrostatic measurements on the boundary. These problems are inverse problems for elliptic equations; see [1] , [2] , [3], [4] , [9], [12] , and their references.

The outline of the paper is as follows. In section 2 we begin introducing the main notations and specifying the a priori information on D and Ω; then we prove the stability result. Section 3 is an appendix where we prove some estimates used in section 2; in particular, we find (Lemma 3.1) an estimate on the analytic continuation in the space variable for solutions of the heat equation. Moreover, we prove (Lemma 3.2) an estimate from below on ∂D for positive solutions of equation (1) fulfilling condition (3).

2. Stability results. Let $x^0 \in \mathbb{R}^3$, $r > 0, A, B$ be two measurable sets in $\mathbb{R}^3$. We use the following notation:

(i) $\mathrm{B}\left(x^0, r\right) = \left\{x \in \mathbb{R}^3 \colon \left|x - x^0\right| < r\right\}$;

(ii) $|A|$ denotes the volume of A; if ∂A is smooth, $|\partial A|$ denotes the area of surface ∂A;

(iii) $\mathrm{d}\left(x^0, A\right) = \inf\limits_{x \in A} \left|x - x^0\right|$, $\mathrm{d}(A, B) = \inf\limits_{x \in A} \mathrm{d}(x, B)$;

(iv) $[A]_r = \left\{x \in \mathbb{R}^3 \colon \mathrm{d}\left(x, A\right) \leq r\right\}$; we set $[A]_0 = \overline{A}$ ($\overline{A}$ is the closure of A);

(v) $(A)_r = \left\{x \in \mathbb{R}^3 \colon \mathrm{d}\left(x, \mathbb{R}^3 \backslash A\right) \geq r\right\}$; we set $(A)_0 = A$;

(vi) $\delta_{\mathcal{H}}\left(A, B\right)$ denotes the Hausdorff distance between A and B;

(vii) $\mathrm{B}'\left(0, r\right) = \left\{x \in \mathbb{R}^2 : |x| < r\right\}$ and

$\mathrm{C}\left(0, r\right) = \left\{x \in \mathbb{R}^3 : \left(x_1, x_2\right) \in \mathrm{B}'\left(0, r\right),\ \left|x_3\right| < r\right\}$;

(viii) $K(x,t) = \begin{cases} \left(2\sqrt{\pi t}\right)^{-3} \exp\left(-\frac{|x|^2}{4t}\right) \text{ if } t > 0,\ x \in \mathbb{R}^3, \\ 0 \text{ if } t \leq 0\ x \in \mathbb{R}^3. \end{cases}$

Let α, d, h, L_0, M, be given positive numbers ($\alpha \in (0,1)$). Ω and D satisfy the a priori information listed in (6), (7).

(6) **Prior information on** Ω :

(a) *Ω is a bounded, $\mathrm{C}^{2,\alpha}$ open set in $\mathbb{R}^3$ whose boundary is connected; d is the diameter of Ω;*

(b) *for every $x \in \partial\Omega$, there exist two balls, $\mathrm{B}\left(x^0, L_0\right)$ $\mathrm{B}\left(y^0, L_0\right)$, such that $\mathrm{B}\left(x^0, L_0\right) \subset \Omega$, $\mathrm{B}\left(y^0, L_0\right) \subset \mathbb{R}^3 \setminus \overline{\Omega,}$ and $\partial\mathrm{B}\left(x^0, L_0\right)$, $\partial\mathrm{B}\left(y^0, L_0\right)$ are tangent to $\partial\Omega$ in x.*

(7) **Prior information on** D :

(a) *D is an open, bounded set in $\mathbb{R}^3$ whose boundary is connected; $\overline{D} \subset \Omega$ and* $\mathrm{d}(D, \partial\Omega) \geq L_0$;

(b) *let $y \in \partial D$ and let $\mathcal{R}$ be the rotation of coordinate that transforms the outer normal to ∂D in y in the vector $(0,0,1)$; then $\mathrm{C}(0,h) \cap \mathcal{R}\left(\partial D - y\right)$ is the graph of a function $g \in \mathrm{C}^{2,\alpha}\left(\mathrm{B}'\left(0,h\right)\right)$ with $g\left(0\right) = 0$, $\nabla g\left(0\right) = 0$, and*

(b_1) $\mathrm{C}(0,h) \cap \mathcal{R}\left(\partial D - y\right) = \left\{x \in \mathbb{R}^3 : \left(x_1, x_2\right) \in \mathrm{B}'\left(0, h\right),\ -h < x_3 < g\left(x_1, x_2\right)\right\}$,

(b_2) $\sup\limits_{\mathrm{B}'(0,h)} \left|g_{x_i x_j}\right| + h^\alpha \sup\limits_{\substack{s,t \in \mathrm{B}'(0,h) \\ s \neq t}} \frac{\left|g_{x_i x_j}(s) - g_{x_i x_j}(t)\right|}{|s-t|^\alpha} \leq M$ *for $i, j = 1, 2$.*

Note that, by denoting $L_1 = \min\left\{h, \frac{L_0}{2}, (2M)^{-1}\right\}$, the following interior and exterior sphere property hold.

Proposition 2.1. *If $r \in [0, L_1)$, then for every $x \in \partial\left[D\right]_r$ there exist two balls $B\left(x^0, L_1\right)$ and $B\left(y^0, L_1 - r\right)$ such that $B\left(x^0, L_1\right) \subset \left[D\right]_r$, $B\left(y^0, L_1 - r\right) \subset \Omega \setminus \left[D\right]_r$, and $\partial B\left(x^0, L_1\right)$, $\partial B\left(y^0, L_1 - r\right)$ are tangent to $\partial\left[D\right]_r$ in x.*

Furthermore, we have the following proposition.

Proposition 2.2. *If $r \in [0, L_1)$, then for every $x \in \Omega \setminus \overline{D}$ such that $d(x, \partial D) = r$ we have $x \in \partial\left[D\right]_r$.*

Proposition 2.3. *If $r_1, r_2 \in [0, L_1)$, then $\left(\Omega\right)_{r_1} \setminus \left[D\right]_{r_2}$ is a connected set.*

The proofs of Propositions 2.1, 2.2, and 2.3 are elementary.

Now we prove the main theorem.

THEOREM 2.4. *Let t_0, t_1 be positive numbers, and let $t_1 > t_0$. Let Ω, D_1, D_2, satisfy, respectively,* (6) *and* (7). *Let $\Sigma \subset \partial\Omega$, Σ nonempty open set of $\partial\Omega$, and $\varphi \in C^1(\partial\Omega \times [0, t_1])$, which fulfills*

$$\varphi(x, 0) = 0 \text{ for } x \in \partial\Omega \tag{8}$$

and

$$\frac{\partial\varphi}{\partial t}(x, t) > 0 \text{ for } (x, t) \in \partial\Omega \times [0, t_1], \tag{9}$$

denote $m_2 = \int\limits_{\partial\Omega} \varphi(x, t_0)\, ds$ by $m_1 = \max\limits_{\partial\Omega\times[0,t_1]} \varphi$.

Let u_j, $j = 1, 2$ satisfy

$$\frac{\partial u_j}{\partial t} - \Delta u_j = 0 \text{ in } \left(\Omega \setminus \overline{D}_j\right) \times [0, t_1], \tag{10}$$

$$u_j = \varphi \text{ on } \partial\Omega \times [0, t_1], \tag{11}$$

$$\frac{\partial u_j}{\partial \mathbf{n}} = 0 \text{ on } \partial D_j \times [0, t_1], \tag{12}$$

$$u_j(x, 0) = 0 \text{ for } x \in \Omega \setminus \overline{D}_j. \tag{13}$$

If

$$\sup_{\Sigma\times[t_0,t_1]} \left| \frac{\partial u_1}{\partial \mathbf{n}} - \frac{\partial u_2}{\partial \mathbf{n}} \right| \le \frac{m_1}{d}\varepsilon, \tag{14}$$

then there exists a nondimensional constant C, depending on $\alpha, d, h, L_0, \frac{m_2}{m_1}, M, t_0, t_1, \Sigma$, and Ω, such that

$$\delta_{\mathcal{H}}(D_1, D_2) \le Cd\left(|\ln\varepsilon|^{-\frac{1}{44}} + \varepsilon\right). \tag{15}$$

Proof. First we normalize the solutions of (10), replacing u_j with $\frac{u_j}{m_1}$ (we continue to denote $\frac{u_j}{m_1}$ with u_j).

Denote

$$u = u_1 - u_2 \text{ in } (\Omega \setminus (D_1 \cup D_2)) \times [0, t_1],$$

$$G = \left(\Omega \setminus (\Omega)_{\frac{L_0}{2}}\right) \times [0, t_1],$$

$$Q_j = \left((\Omega)_{\frac{L_0}{2}} \setminus D_j\right) \times [0, t_1].$$

By regularity estimates for parabolic equations, we have

$$\|u\|_{L^\infty(G)} + d\|\nabla_x u\|_{L^\infty(G)} + d^2\left\|D_x^2 u\right\|_{L^\infty(G)} \le E_1, \tag{16}$$

where E_1 is a nondimensional constant depending on α, L_0, Ω, and

$$\|u_j\|_{L^\infty(Q_j)} + d\|\nabla_x u_j\|_{L^\infty(Q_j)} + d^2\left\|D_x^2 u_j\right\|_{L^\infty(Q_j)} + d^3\|\nabla_x D_t u_j\|_{L^\infty(Q_j)} \le E_2, \tag{17}$$

where E_2 depends on α, d, h, L_0, M(in 16); in (17) D_x^2 denotes the matrix of second derivatives with respect to x_i, x_k). We set

$$E = \max\{E_1, E_2\}.$$

Starting from the stability estimates for the Cauchy problem for parabolic equations contained in [13, Chap. IV], estimating the constants appearing there and applying Lemma 3.1 to u_{x_k}, $k = 1, 2, 3$ yields

$$|\nabla_x u(x, t_2)| \le \frac{C_{11}}{d}(E_1 + \varepsilon)^{1-\gamma_1} \varepsilon^{\gamma_1} \text{ for } \frac{L_0}{8} \le \mathrm{d}(x, \partial\Omega) \le \frac{L_0}{2}, \tag{18}$$

where

$$t_2 = \frac{t_0 + t_1}{2},$$

$$C_{11} = C_{10}\left(\frac{\tau}{dL_0} + \frac{d^4}{\tau^{\frac{3}{2}} L_0}\right),$$

$$\tau = t_1 - t_0,$$

$$\gamma_1 = \exp - c\left(C'_{10}\left(dL_0^{-1} + d\tau^{-\frac{1}{2}}\right) + |\Omega| L_0^{-3}\right)$$

(C_{10} depends only on Σ, d, t_1; C'_{10} depends on Σ and d; and c is an absolute constant).

The proof is subdivided into two steps. In the first we prove a log–log estimate for $\delta_{\mathcal{H}}(D_1, D_2)$; in the second we prove estimate (15).

Step 1. Let $r \in \left(0, \frac{L_1}{4}\right)$. Denote by I_r the connected component of $(\Omega)_r \setminus ([D_1]_r \cup [D_2]_r)$ that contains $\partial(\Omega)_r$. By positivity of $\frac{\partial u_2}{\partial t}(x, t_2)$ in $\Omega \setminus D_2$, the inclusion $D_1 \setminus D_2 \subset (((\Omega)_r \setminus I_r) \setminus [D_2]_r) \cup ([D_2]_r \setminus D_2)$, estimate (17), and inequality (69), we have

$$\int\limits_{D_1 \setminus D_2} \frac{\partial u_2(x, t_2)}{\partial t} dx \le \int\limits_{[D_2]_r \setminus D_2} \frac{\partial u_2(x, t_2)}{\partial t} dx + \int\limits_{\left((\Omega)_r \setminus I_r\right) \setminus [D_2]_r} \frac{\partial u_2(x, t_2)}{\partial t} dx \tag{19}$$

$$\le c\frac{E_2 r d}{M L_1^3} + \int\limits_{\left((\Omega)_r \setminus I_r\right) \setminus [D_2]_r} \Delta u_2(x, t_2)\, dx.$$

By (10), (12), the divergence theorem (see [7]), the inclusion

$$\partial(((\Omega)_r \setminus I_r) \setminus [D_2]_r) \subset (\partial I_r \cap \partial [D_1]_r) \cup \partial [D_2]_r, \tag{20}$$

estimate (17), and inequality (69), we obtain

$$\int\limits_{\left((\Omega)_r \setminus I_r\right) \setminus [D_2]_r} \Delta u_2(x, t_2)\, dx \le \int\limits_{\left(\partial I_r \cap \partial [D_1]_r\right) \cup \partial [D_2]_r} \left|\frac{\partial u_2(x, t_2)}{\partial \mathbf{n}}\right| ds \tag{21}$$

$$\leq \int_{\partial I_r \cap \partial [D_1]_r} |\nabla_x u(x,t_2)|\, ds + \int_{\partial [D_2]_r} \left| \frac{\partial u_2(x,t_2)}{\partial \mathbf{n}} \right| ds + \int_{\partial [D_1]_r} \left| \frac{\partial u_1(x,t_2)}{\partial \mathbf{n}} \right| ds$$

$$\leq \frac{cd^3}{M^2 L_1^3} \left(\max_{x \in \partial I_r} |\nabla_x u(x,t_2)| + \frac{E_2 r}{d^2} \right).$$

From estimates (19) and (21) and applying inequality (66) (see Remark 3 of section 3), where u is replaced with $\frac{\partial u_2}{\partial t}$, T with t_2, T'' with t_0, and T' with $\frac{t_0+t_2}{2}$, we have

$$|D_1 \setminus D_2| \leq C_{12} d^2 \left(E_2 r + d^2 \max_{x \in \partial I_r} |\nabla_x u(x,t_2)| \right), \tag{22}$$

where C_{12} depends on $\alpha, d, h, L_0, \frac{m_2}{m_1}, M$, and τ.

Now consider the set $J = \left\{ x \in (\Omega)_r : \mathrm{d}(x, \partial\Omega) \leq \frac{L_0}{2} \right\}$. By the interior sphere property of Ω, it follows that J is connected. Furthermore, J is contained in $(\Omega)_r \setminus ([D_1]_r \cup [D_2]_r)$ and $\partial(\Omega)_r \subset J$; therefore, the inclusion $J \subset I_r$ is valid. Let $\overline{x} \in \partial I_r$, $y \in J$ such that $\mathrm{d}(y, \partial\Omega) = \frac{L_0}{4}$, and Γ be a simple curve in I_r with endpoints $\overline{x}$ and y. Γ is contained in $(\Omega)_r \setminus (D_1 \cup D_2)$, and $\mathrm{d}(\Gamma, D_j) \geq r$, $j = 1, 2$, $\mathrm{d}(\Gamma, \partial\Omega) \geq r$. We construct a chain of closed balls which are centered in Γ, have radius $\frac{r}{3}$, are tangent two by two, and are internally nonoverlapping. The first ball is centered in y; the last is at a distance less than or equal to $\frac{r}{3}$ from $\overline{x}$. If the distance is less than $\frac{r}{3}$, we add to the chain the ball of radius $\frac{r}{3}$ whose center is $\overline{x}$. The repeated use of Lemma 3.1 applied to u_{x_j}, $j = 1, 2, 3$ gives

$$|\nabla_x u(\overline{x}, t_2)| \leq 6 (1 + C_0)^{2N_r} \left(E d^{-1} + \varepsilon_1 \right)^{1 - \gamma_0^{2N_r}} \varepsilon_1^{\gamma_0^{2N_r}}, \tag{23}$$

where $C_0 = e^{\frac{L_1^2}{64 t_2}}$ and γ_0 is the exponent in the inequality (42), ε_1 is defined by

$$\varepsilon_1 = \max_{\mathrm{B}\left(y, \frac{r}{3}\right)} |\nabla_x u(x,t_2)|,$$

and

$$N_r = \frac{3^4 d^3}{4\pi r^3}.$$

By estimates (18) and (23) we obtain

$$|D_1 \setminus D_2| \leq C_{13} d^2 (E + \varepsilon) \left(r + d (1 + C_0)^{2N_r} \left(\frac{\varepsilon}{E + \varepsilon} \right)^{\gamma_1 \gamma_0^{2N_r}} \right), \tag{24}$$

where C_{13} depends on $\alpha, d, h, L_0, \frac{m_2}{m_1}, M, \tau, t_1$, and Σ.

Denoting by l the integral part of $\frac{1}{\gamma_1} \left(1 + \frac{\ln(1+C_0)}{|\ln \gamma_1|} \right)$ and ε_0 the number

$$\varepsilon_0 = E \exp\left(-\left((2l)!\right)^2 \exp\left(\frac{2 \cdot 3^4 d \left(|\ln \gamma_0| + 1 \right)}{\pi L_1} \right) \right)$$

for $\varepsilon \leq \varepsilon_0$, we can choose in (24) r given by

$$r = \frac{L_1}{4} \left(\frac{3^4 |\ln \gamma_1| \left(4 d L_1^{-1} \right)^3}{4\pi \ln \left(\left| \ln \frac{\varepsilon}{E+\varepsilon} \right| \right)^{\frac{1}{2}}} \right)^{\frac{1}{3}},$$

and we have

$$|D_1 \setminus D_2| \leq C_{14} d^3 \left(\ln \left|\ln \frac{\varepsilon}{E+\varepsilon}\right|\right)^{-\frac{1}{3}}, \tag{25}$$

where C_{14} depends on $\alpha, d, h, L_0, \frac{m_2}{m_1}, M, \tau, t_1$, Σ, and Ω. The same inequality holds for $|D_2 \setminus D_1|$.

Denote by $\sigma(\varepsilon)$ the number

$$\sigma(\varepsilon) = C_{14} \left(\ln \left|\ln \frac{\varepsilon}{E+\varepsilon}\right|\right)^{-\frac{1}{3}}$$

(here C_{14} is the same constant of estimate (25)). By Proposition 3.5 we have

$$\delta_{\mathcal{H}}(D_1, D_2) \leq 2\left(d^7 L_1^{-3} \sigma(\varepsilon)\right)^{\frac{1}{4}}. \tag{26}$$

Step 2. Let I_0 denote the connected component of $\overline{\Omega} \setminus (D_1 \cup D_2)$ that contains $\partial\Omega$. By (10), (12), inclusion (20) considered for $r = 0$, and the divergence theorem we have

$$\int\limits_{D_1 \setminus D_2} \frac{\partial u_2(x, t_2)}{\partial t} dx \leq \int\limits_{(\partial I_0 \cap \partial D_1) \setminus \partial D_2} |\nabla_x u(x, t_2)| \, ds. \tag{27}$$

Let $\overline{x} \in (\partial I_0 \cap \partial D_1) \setminus \partial D_2$. By (26), we have

$$\mathrm{d}(\overline{x}, D_2) \leq \delta(\varepsilon), \tag{28}$$

where $\delta(\varepsilon) = 2\left(d^7 L_1^{-3} \sigma(\varepsilon)\right)^{\frac{1}{4}}$.

The definition of I_0 states that $\overline{x} \in \Omega \setminus \overline{D}_2$. Let $r = d(\overline{x}_1, D_2)$. Furthermore, let ε_1 be such that $\varepsilon_1 \leq \varepsilon_0$ and $\delta(\varepsilon_1) \leq \frac{L_1}{2}$; (28) and Proposition 2.1 yield, for $\varepsilon \leq \varepsilon_1$, that $\overline{x} \in \partial [D_2]_r \cap \partial D_1$. By Proposition 2.2 we have that there exist four balls $\mathrm{B}_i^{(j)}$, $\mathrm{B}_e^{(j)}$, $j = 1, 2$, with radius $\frac{L_1}{2}$, which fulfill

(a) $\mathrm{B}_i^{(1)} \subset D_1$, $\mathrm{B}_e^{(1)} \subset \Omega \setminus \overline{D}_1$,
with $\partial \mathrm{B}_i^{(1)}$ and $\partial \mathrm{B}_e^{(1)}$ tangent to ∂D_1 in $\overline{x}$;

(b) $\mathrm{B}_i^{(2)} \subset [D_2]_r$, $\mathrm{B}_e^{(2)} \subset \Omega \setminus [D_2]_r$,
with $\partial \mathrm{B}_i^{(2)}$ and $\partial \mathrm{B}_e^{(2)}$ tangent to $\partial [D_2]_r$ in $\overline{x}$.

Note that

$$\mathrm{B}_i^{(1)} \cap \mathrm{B}_e^{(2)} \subset D_1 \setminus [D_2]_r, \tag{29}$$

$$\mathrm{B}_e^{(1)} \cap \mathrm{B}_e^{(2)} \subset \Omega \setminus \left(\overline{D}_1 \cup [D_2]_r\right). \tag{30}$$

By (29) and (30) we have

$$\left|\mathrm{B}_e^{(2)} \setminus \mathrm{B}_i^{(1)}\right| \geq \left|\mathrm{B}_e^{(2)}\right| - |D_1 \setminus [D_2]_r| \geq \frac{4\pi}{3}\left(\frac{L_1}{2}\right)^3 - d^3 \sigma(\varepsilon). \tag{31}$$

Now consider the planes $\mathcal{P}_1$ and $\mathcal{P}_2$ which are tangent in $\overline{x}$, respectively, to $\mathrm{B}_i^{(1)}$, $\mathrm{B}_e^{(1)}$ and $\mathrm{B}_i^{(2)}$, $\mathrm{B}_e^{(2)}$. $\mathrm{B}_e^{(1)} \cap \mathrm{B}_e^{(2)}$ is contained in one of the four regions in which $\mathbb{R}^3$ is divided by $\mathcal{P}_1$ and $\mathcal{P}_2$. Let $\mathcal{P}_j^+ \subset \mathcal{P}_j$, $j = 1, 2$, the half-planes that limit this region, and denote by θ, $\theta \in [0, \pi)$ the angle between $\mathcal{P}_1^+$ and $\mathcal{P}_2^+$.

By the formula

$$\left|\mathrm{B}_e^{(2)} \setminus \mathrm{B}_i^{(1)}\right| = 2\pi \left(\frac{L_1}{2}\right)^3 \left(\sin\frac{\theta}{2} - \frac{1}{3}\sin^3\frac{\theta}{2}\right) \tag{32}$$

and by (31), we have, if $\varepsilon \leq \varepsilon_2$ (where ε_2 is such that $\sigma(\varepsilon_2) \leq \frac{\pi}{100}\left(\frac{L_1}{2}\right)^3$ and $\varepsilon_2 \leq \varepsilon_1$),

$$\sin\frac{\theta}{2} - \frac{1}{3}\sin^3\frac{\theta}{2} \geq \frac{3\sqrt{3}}{8}\ .$$

Therefore, $\theta \geq \theta_0 > \frac{3}{4}\pi$. Denote $S_\theta(\overline{x})$ $=\mathrm{B}_e^{(1)}\cap\mathrm{B}_e^{(2)}$; let $\mathcal{L}$ be the plane containing the circle $\partial\mathrm{B}_e^{(1)} \cap \partial\mathrm{B}_e^{(2)}$. We can suppose, by mean of an isometry, that the center of symmetry of $S_\theta(\overline{x})$ is in the origin of coordinate, $\mathcal{L}$ is the plane $x_3 = 0$, and $\overline{x}$ belongs to x_1 axis and $\overline{x}_1 > 0$. Denote $s_0 = \frac{L_1}{8}\left(1 - \cos\frac{\theta_0}{2}\right)$. We have $\mathrm{B}(0, s_0) \subset S_\theta(\overline{x})$ and $\mathrm{d}(\mathrm{B}(0, s_0), D_j) \geq 2s_0$, $j = 1, 2$. Let $\varepsilon \leq \varepsilon_3$, where ε_3 is such that $\delta(\varepsilon_3) \leq \frac{s_0}{2}$ and $\varepsilon_3 \leq \varepsilon_2$. By (26) we have $D_1 \cup D_2 \subset [D_1]_{\frac{s_0}{2}} \subset [D_2]_{s_0}$; therefore, $\mathrm{B}(0, s_0) \subset (\Omega)_{s_0} \setminus [D_2]_{s_0}$. By Proposition 2.3, $(\Omega)_{s_0} \setminus [D_2]_{s_0}$ is connected; hence we can join every point $x \in$B$(0, s_0)$ with $y \in \Omega$ such that $\mathrm{d}(y, \partial\Omega) = \frac{L_0}{4}$. Proceeding as in the proof of (23), we obtain, by (18),

$$|\nabla_x u(x, t_2)| \leq \frac{\omega(\varepsilon)}{d} \text{ for every } x \in \mathrm{B}(0, s_0), \tag{33}$$

where

$$\omega(\varepsilon) = 12(1 + C_0)^{N_0} C_{10} (E + \varepsilon)^{1-\gamma_1\gamma_0^{2N_0}} \varepsilon^{\gamma_1\gamma_0^{2N_0}}$$

and

$$N_0 = \frac{3^4 d^3}{4\pi s_0^3}.$$

Let $\rho \in \left(0, \frac{L_1}{4}\sin\frac{\theta_0}{2}\right)$ and denote

$$x_\rho = (\overline{x}_1 - 2\rho, 0, 0),\ r = \frac{1}{2}\mathrm{d}(x_\rho, \partial S_\theta(\overline{x})).$$

Observe that

$$\frac{1}{2}\rho\sin\frac{\theta}{2} < r < \rho. \tag{34}$$

Using the Green formula for u_{x_j} in $S_\theta(\overline{x}) \times [0, t_2]$ for fix $j \in \{1, 2, 3\}$ and setting

$$f(\lambda + i\mu) = u_{x_j}(\lambda + i\mu, 0, 0, t_2), \tag{35}$$

a straightforward calculation shows that the complex analytic function f is defined in the subset of $\mathbb{C}$:

$$\Lambda = \left\{\lambda + i\mu : -s_0 \leq \lambda \leq \overline{x}_1 - 2\rho + r, |\mu| \leq \frac{1}{2}\left(\frac{L_1}{2} - \sqrt{\lambda^2 + \frac{L_1^2\cos^2\frac{\theta}{2}}{4}}\right)\right\}. \tag{36}$$

Furthermore, by (33), the following estimates follow:

$$\left|f\left(\lambda+i0\right)\right|\leq\frac{\omega\left(\varepsilon\right)}{d}\text{ for }\left|\lambda\right|\leq s_0,$$

$$\left|f\left(\lambda+i\mu\right)\right|\leq\frac{cEL_1^3e^{\frac{L_1^2}{64T}}}{r^3d}\text{ for }\lambda+i\mu\in\Lambda,$$

where c is an absolute constant.

These inequalities and estimates for analytic continuation, see [5] , give (recalling that $u_{x_j}\left(x_\rho,t_2\right)=f\left(\overline{x}_1-2\rho+i0\right)$)

$$\left|u_{x_j}\left(x_\rho,t_2\right)\right|\leq\frac{cL_1^3e^{\frac{L_1^2}{64T}}}{r^3d}\left(E+\omega\left(\varepsilon\right)\right)^{1-A_r}\left(\omega\left(\varepsilon\right)\right)^{A_r},\tag{37}$$

where

$$A_r=c_1\left(\frac{r}{L_1}\right)^{10}$$

and c, c_1 are absolute constants.

From Proposition 2.1, (17), (27), (37), and (67), we have

$$\left|D_1\setminus D_2\right|\leq C_{15}d^2\left(E+\varepsilon\right)\left(\rho+\frac{d^4}{r^3}\left(\frac{\varepsilon}{E+\varepsilon}\right)^{\gamma_2A_r}\right),\tag{38}$$

where $\gamma_2=\gamma_1\gamma_0^{2N_0}$ and C_{15} depends on $\alpha,d,h,L_0,\frac{m_2}{m_1},M,\tau,t_1,\Sigma,$ and Ω. If $\varepsilon\leq\varepsilon_4$, where $\varepsilon_4=\min\{\varepsilon_3,E\exp(-\exp\frac{2^{140}}{\gamma_2C_1})\}$, and we set in (38)

$$\rho=L_1\left|\ln\frac{\varepsilon}{E}\right|^{-\frac{1}{11}},$$

we have, recalling (34),

$$\left|D_1\setminus D_2\right|\leq C_{16}Ed^3\left|\ln\frac{\varepsilon}{E}\right|^{-\frac{1}{11}},$$

where C_{16} depends on the same data of C_{15}. The same estimate holds for $\left|D_2\setminus D_1\right|$.

Now, proceeding as in the proof of (26), we obtain

$$\delta_{\mathcal{H}}\left(D_1,D_2\right)\leq 2d\left(C_{16}E\right)^{\frac{1}{4}}\left|\ln\frac{\varepsilon}{E}\right|^{-\frac{1}{44}}.$$

Finally, by the trivial inequality $\delta_{\mathcal{H}}\left(D_1,D_2\right)\leq\frac{d}{\varepsilon_4}\varepsilon$ valid for $\varepsilon>\varepsilon_4$, we obtain (15).

Remark 1. More careful and tedious calculations in the evaluation of angle θ and of analytic continuation of f in Λ give an improvement, only concerning the epsilon dependence, in estimate (15). More precisely, (i) by inequality (31) and formula (32) we have that for every $\nu>0$ there exists $\omega_\nu>0$ ($\omega_\nu\rightarrow 0$ as $\nu\rightarrow 0$) such that if $\varepsilon\leq\omega_\nu$ then $\theta\geq\pi-\nu$; (ii) to estimate $f\left(\lambda+i\mu\right)$ in Λ we can use a more large part than really was used in the proof—in this way the exponent γ_2A_r in (38) can be replaced by $\varkappa_\nu\left(\frac{r}{R}\right)^{2+\varkappa_\nu}$, where $\varkappa_\nu\rightarrow 0$ as $\nu\rightarrow 0$, the constant C_{15} can be replaced by C_ν ($C_\nu \to \infty$ as $\nu \to 0$), and if we choose $\rho = L_1 \left|\ln \frac{\varepsilon}{E}\right|^{-\left(\frac{1+\omega_\nu}{2+\omega_\nu}\right)}$, we obtain the estimate

$$\delta_{\mathcal{H}}\left(D_1, D_2\right) = O\left(\left|\ln\varepsilon\right|^{-\frac{1}{8}+\omega_\nu}\right). \tag{39}$$

Remark 2. If in the a priori information on D, instead of ∂D, connected, we assume that $\mathbb{R}^3 \setminus \overline{D}$ is connected, $D = \bigcup_{n=1}^{N} D^{(n)}$ with $\overline{D^{(i)}} \cap \overline{D^{(j)}} = \emptyset$ for $i \neq j$, and $\partial D^{(i)}$ is connected for every $i = 1, ..., N$, then estimate (15) (and (39)) continues to hold with a different constant.

3. Appendix. In the following lemma we prove an estimate on the analytic continuation of the space variable for solutions of the heat equation.

LEMMA 3.1. *Let $\varepsilon, E, r, T > 0$. Suppose that u satisfies the equation*

$$\frac{\partial u}{\partial t} - \Delta u = 0 \textit{ in } B(0, 3r) \times [0, T].$$

Let $y \in \partial B(0, r)$. Suppose that

$$\|u\|_{L^\infty(B(0,3r)\times[0,T])} + r\left\|\nabla_x u\right\|_{L^\infty(B(0,3r)\times[0,T])} \leq E, \tag{40}$$

$$\left|u\left(x, T\right)\right| \leq \varepsilon \textit{ for } x \in B\left(y, r\right) \cap B\left(0, r\right). \tag{41}$$

The following estimate holds:

$$\left|u\left(x, T\right)\right| \leq C_0\left(E + \varepsilon\right)^{1-\gamma_0} \varepsilon^{\gamma_0} \textit{ for } x \in B\left(y, r\right) \setminus B\left(0, r\right), \tag{42}$$

where $C_0 = 240 e^{\frac{r^2}{64T}}$ and $\gamma_0 = \frac{1}{4\cosh\frac{\pi}{8}\cosh\pi}$.

Proof. The Green formula gives, for $x \in$B$(0, 3r)$,

$$u\left(x, T\right) = \int\limits_{\mathrm{B}(0,3r)} K\left(x - \xi, T\right) u\left(\xi, 0\right) d\xi \tag{43}$$

$$+ \int_0^T dt \int\limits_{\partial\mathrm{B}(0,3r)} \left(K\left(x - \xi, T - t\right)\frac{\partial u\left(\xi, t\right)}{\partial \mathbf{n}_\xi} - \frac{\partial K}{\partial \mathbf{n}_\xi}\left(x - \xi, T - t\right) u\left(\xi, t\right)\right) ds_\xi.$$

Now, let $z_k = x_k + iy_k, k = 1, 2, 3$, and consider the function $U\left(z\right)$ obtained by the formal substitution in the right-hand side of (43) of $x_k$with z_k. $U\left(z\right)$ is the complex analytic continuation of $u\left(x, T\right)$; in fact, $U\left(x_1 + i0, x_2 + i0, x_3 + i0\right) = u\left(x, T\right)$ and a simple calculation prove that for $x \in$B$(0, 3r)$, $y \in$B$(0, \frac{3r-|x|}{2})$, the Cauchy–Riemann equations $\frac{\partial U}{\partial x_k} + i\frac{\partial U}{\partial y_k} = 0$, $k = 1, 2, 3$, are valid. We continue to denote by $u\left(z, T\right)$ the function $U\left(z\right)$.

Consider $\overline{x} \in$B$(y, r) \setminus$B$(0, r)$ and denote

$$f\left(\lambda + i\mu\right) = u\left(\left(\lambda + i\mu\right)\overline{x}, T\right).$$

From (40) and (43) we have

$$\left|f\left(\lambda + i\mu\right)\right| \leq 120 E e^{\frac{r^2}{64T}} \text{ for } \lambda \in \left[0, \frac{5r}{2\left|\overline{x}\right|}\right], \mu \in \left[-\frac{r}{4\left|\overline{x}\right|}, \frac{r}{4\left|\overline{x}\right|}\right]. \tag{44}$$

Furthermore, by (41), we obtain

$$|f(\lambda+i0)| \leq \varepsilon \text{ for } \lambda \in \left[0, \frac{r}{|\overline{x}|}\right]. \tag{45}$$

By analytic continuation in the rectangle $[0, \frac{5r}{|\overline{x}|}] \times [-\frac{r}{4|\overline{x}|}, \frac{r}{4|\overline{x}|}]$ (see [5]), we have by (44) and (45) that

$$|u(\overline{x},T)| = |f(1+i0)| \leq 240 e^{\frac{r^2}{64T}} (E+\varepsilon)^{1-\gamma_0} \varepsilon^{\gamma_0},$$

where $\gamma_0 = \frac{1}{4\cosh\frac{\pi}{8}\cosh\pi}$. □

LEMMA 3.2. *Let Ω and D satisfy* (6) *and* (7), *respectively. Let u be a solution of the equation*

$$\frac{\partial u}{\partial t} - \Delta u = 0 \textit{ in } \left(\Omega \setminus \overline{D}\right) \times [0,T],$$

which is positive in $\left(\Omega \setminus \overline{D}\right) \times (0,T]$ and fulfills the condition

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \textit{ on } \partial D \times [0,T].$$

The following estimate then holds:

$$\min_{x\in\partial D} u(x,T) \geq \frac{C}{L_0^4} \int_0^{T'} dt \int_{\partial\Omega} u(x,t)\, ds, \tag{46}$$

where $T' \in (0,T)$ and C is a nondimensional constant depending on $d, h, L_0, M, T, T-T'$.

The proof of this lemma is preceded by the following.

LEMMA 3.3. *Let Ω and D be as in the previous lemma. Let $x^0 \in \partial D$, $\sigma > 0$, and let v be the solution of*

$$\frac{\partial v}{\partial t} - \Delta v = 0 \textit{ in } \left(\Omega \setminus \overline{D}\right) \times [0,T,], \tag{47}$$

$$v(x,0) = K\left(x-x^0,\sigma\right) \textit{ for } x \in \Omega \setminus \overline{D}, \tag{48}$$

$$v = 0 \textit{ on } \partial\Omega \times [0,T], \tag{49}$$

$$\frac{\partial v}{\partial \mathbf{n}} = 0 \textit{ on } \partial D \times [0,T]. \tag{50}$$

For every $T' \in (0,T)$ there exist $\sigma_0 > 0$ and C_8 depending on $d, h, L_0, L_1, M, T, T-T'$ such that for $\sigma \in (0,\sigma_0)$ and $(x,t) \in \partial\Omega \times [T-T', T]$,

$$-\frac{\partial v(x,t)}{\partial \mathbf{n}} \geq \frac{1}{L_0^4} \exp -\frac{c}{C_8}\left[\left(dL_1^{-1}\right)^6 + TL_1^{-2}\right], \tag{51}$$

where c is an absolute constant and $\mathbf{n}$ is the outward normal to $\partial\Omega$.

Proof of Lemma 3. First we find an estimate from below in a point far enough from D for v. We construct v as follows: $v = w + z$, where

$$z\left(x,t\right) = K\left(x - x^0, t + \sigma\right) + \int_0^t d\eta \int_{\partial D} K\left(x - \xi, t - \eta\right) \psi\left(\xi, \eta\right) d\xi, \tag{52}$$

$\psi \in \mathrm{C}^0\left(\partial D \times [0,T]\right)$, and it is such that

$$\frac{\partial z}{\partial \mathbf{n}} = 0 \text{ on } \partial D \times [0,T]. \tag{53}$$

w is the solution of the boundary value problem (the existence of w is assured by [14, Chap. III]);

$$\frac{\partial w}{\partial t} - \Delta w = 0 \text{ in } \left(\Omega \setminus \overline{D}\right) \times [0,T],$$

$$w\left(x,0\right) = 0 \text{ for } x \in \Omega \setminus \overline{D},$$

$$w = -z \text{ on } \partial\Omega \times [0,T],$$

$$\frac{\partial w}{\partial \mathbf{n}} = 0 \text{ on } \partial D \times [0,T].$$

Condition (53) is equivalent to the integral equation

$$\psi\left(x,t\right) - \int_0^t d\eta \int_{\partial D} G\left(x,t;\xi,\eta\right) \psi\left(\xi,\eta\right) ds_\xi = F_0\left(x,t\right) \text{ for } \left(x,t\right) \in \partial D \times [0,T],$$

where

$$G\left(x,t;\xi,\eta\right) = 2\frac{\partial K}{\partial \mathbf{n}_x}\left(x - \xi, t - \eta\right) \text{ for } 0<\eta<t,\ x,\xi \in \partial D,$$

$$F_0\left(x,t\right) = -2\frac{\partial K}{\partial \mathbf{n}_x}\left(x - x^0, t + \sigma\right) \text{ for } \left(x,t\right) \in \partial D \times [0,T],$$

and $\mathbf{n}_x$ denotes unit inner normal to ∂D in x.

The continuity of F_0 gives the continuity of ψ (see [8]); furthermore,

$$\psi\left(x,t\right) = \sum_{m=0}^{\infty} F_m\left(x,t\right), \tag{54}$$

where

$$F_m\left(x,t\right) = \int_0^t d\eta \int_{\partial D} G\left(x,t;\xi,\eta\right) F_{m-1}\left(\xi,\eta\right) ds_\xi \tag{55}$$

for $m \geq 1$, $(x,t) \in \partial D \times [0,T]$.

Now for $\mu \in (0,1)$ we have

$$|G(x,t;\xi,\eta)| \leq \frac{c\left(M+h^{-1}\right)}{(t-\eta)^{\mu}\,|x-\xi|^{3-2\mu}}, \tag{56}$$

$$|F_0(x,t)| \leq \frac{c\left(M+h^{-1}\right)}{(t+\sigma)^{\mu}\,|x-x^0|^{3-2\mu}}, \tag{57}$$

where c is a constant depending on μ.

Furthermore, by a slight modification of [8, Lemma 1, p. 137], we have for $x \in \partial D$, $\xi \in \Omega \setminus D$, and $a, b \in [0,2)$,

$$\int\limits_{\partial D} \frac{ds_y}{|x-y|^{a}\,|y-\xi|^{b}} \leq \begin{cases} \frac{C_1}{|x-\xi|^{a+b-2}} \text{ if } a+b>2, \\ \\ C_2 \text{ if } a+b<2, \end{cases} \tag{58}$$

where C_1 and C_2 are constants depending on a, b, d, h, M.

Let $\mu = \frac{1}{\sqrt{3}}$ (observe that $\mu \in \left(\frac{1}{2}, \frac{2}{3}\right)$; by (55)–(58) we have

$$|F_m(x,t)| \leq \begin{cases} \frac{C_3}{d}\frac{\left(C_3 t^{1-\mu} d^{-1}|x-x^0|^{2\mu-1}\right)^m t^{-\mu}}{\Gamma((m+1)(1-\mu))|x-x^0|^{3-2\mu}} \text{ for } 2 \leq m \leq N, \\ \\ \frac{C_3}{d^{4-2\mu}}\frac{\left(C_3 t^{1-\mu} d^{2(\mu-1)}\right)^m t^{-\mu}}{\Gamma((m+1)(1-\mu))} \text{ for } m \geq N+1, \end{cases} \tag{59}$$

where N is the integral part of $\frac{3-2\mu}{2\mu-1}$ and C_3 is a constant depending on d, h, M.

Furthermore,

$$|F_0(x,t)| + |F_1(x,t)| \leq \frac{C_4}{d^{1+2\mu}}\left(\frac{T}{t+\sigma}\right)^{\mu}\frac{1}{|x-x^0|^{3-2\mu}}\ , \tag{60}$$

where C_4 depends on d, h, M, T.

By (54), (59), (60) we obtain

$$|\psi(x,t)| \leq \frac{C_4'}{d^{1+2\mu}}\left(\frac{T}{t+\sigma}\right)^{\mu}\frac{1}{|x-x^0|^{3-2\mu}}\ ; \tag{61}$$

C_4' depends on d, h, M, T.

This inequality, (52), and the maximum principle yield

$$\|w\|_{L^{\infty}((\Omega\setminus D)\times[0,T])} \leq \|z\|_{L^{\infty}(\partial\Omega\times[0,T])} \leq \frac{C_4''}{L_0^3}; \tag{62}$$

C_4'' depends on d, h, M, T.

From (52), (61), and (62) we obtain

$$v\left(x, |x-x^0|^2-\sigma\right) \geq \frac{1}{|x-x^0|^3}\left[\frac{1}{e^{1/4}\left(2\sqrt{\pi}\right)^3} - \frac{C_5\,|x-x^0|}{L_0}\right], \tag{63}$$

where C_5 depends on d, h, L_0, M, T. For such a choice of C_5, denote

$$C_6 = \min\left\{\left(2e^{1/4}\left(2\sqrt{\pi}\right)^3 C_5\right)^{-1}, \frac{L_1}{2L_0}\right\},\ C_7 = \left(2e^{1/4}\left(2\sqrt{\pi}\right)^3 C_6^3\right)^{-1}.$$

We obtain, for $x \in \left(\Omega \setminus \overline{D}\right) \cap \partial \mathrm{B}\left(x^0, C_6 L_0\right)$ and $\sigma < \frac{C_6^2 L_0^2}{2}$, the following estimate:

$$v\left(x, C_6^2 L_0 - \sigma\right) \geq \frac{C_7}{L_0^3}\ . \tag{64}$$

This estimate, the Harnack inequality, and the Hopf maximum principle give, for $T' \in (0,T)$,

$$-\frac{\partial v}{\partial \mathbf{n}}\left(x,t\right) \geq \frac{1}{L_0^4} \exp -\frac{c}{C_8}\left(\left(dL_1^{-1}\right)^6 + TL_1^{-2}\right) \text{ for } (x,t) \in \partial\Omega \times \left[T - T', T\right],$$

where $\mathbf{n}$ is the outward normal to $\partial\Omega$, c is an absolute constant,

$$C_8 = \min\left\{C_7^6, \frac{\left(T - T'\right)^{3/2}}{\left(4L_0\right)^3}\right\},$$

and C_7 is as in (64). □

Proof of Lemma 2. Let $x^0 \in \partial D$ such that

$$u\left(x^0, T\right) = \min_{x \in \partial D} u\left(x, T\right).$$

By the Green formula applied to u and $v\left(x, T - t\right)$, the positivity of u, and (51) we obtain

$$\frac{1}{2} u\left(x^0, T\right) = \lim_{\sigma \to 0^+} \int\limits_{\Omega \setminus D} u\left(x, T\right) K\left(x - x^0, \sigma\right) dx \geq \lim_{\sigma \to 0^+} \int\limits_0^T dt \int\limits_{\partial\Omega} u\left(x,t\right)\left(-\frac{\partial v}{\partial \mathbf{n}}\right)\left(x, T - t\right) ds$$

$$\geq \frac{1}{L_0^4}\left(\exp -\frac{c}{C_8}\left(\left(dL_1^{-1}\right)^6 + TL_1^{-2}\right)\right) \int\limits_0^{T'} dt \int\limits_{\partial\Omega} u\left(x,t\right) ds.$$

That is, we obtain (46). □

Remark 3. If $\|\nabla_x u\|_{L^\infty\left(\left(\Omega\right)_{L_0} \setminus D \times [0,T]\right)} \leq d^{-1}E$, $E > 0$, by (46) we have for $x \in \Omega \setminus D$ such that $\mathrm{d}(x, \partial D) \leq \frac{Cd}{L_0^4 E} \int_0^{T'} dt \int_{\partial\Omega} u ds$

$$u\left(x, T\right) \geq \frac{C}{2L_0^4} \int\limits_0^{T'} dt \int\limits_{\partial\Omega} u\left(x,t\right) ds. \tag{65}$$

Furthermore, if $T'' \in (0, T')$, using (65) where T is replaced by T' and T' by T'' and applying again the Harnack inequality, we get

$$u\left(x, T\right) \geq \frac{C_9}{L_0^4} \int\limits_0^{T''} dt \int\limits_{\partial\Omega} u\left(x,t\right) ds \text{ for } x \in \Omega \setminus D \text{ such that } \mathrm{d}\left(x, \partial\Omega\right) \geq L_0, \tag{66}$$

where C_9 depends on $d, E, h, L_0, M, T, T - T', T' - T''$.

Proposition 3.4. *Let D fulfill condition* (7), *denote* $L_2 = \min\{(2M)^{-1}, h\}$; *we have*

$$|\partial D| \leq \frac{24e^3}{M^2 L_2^3} |D| , \tag{67}$$

$$|\partial [D]_r| \leq \frac{96e^3}{M^2 L_2^3} |D| \ \textit{for } r \in (0, L_2) , \tag{68}$$

$$|[D]_r \setminus D| \leq \frac{56e^3}{M^2 L_2^3} |D| \, r. \tag{69}$$

Proof. Denote by $H(x)$ and $K(x)$, respectively, the mean and total curvature of ∂D in x. By (7) we have

$$|H(x)| \leq \min\left\{M, L_2^{-1}\right\}, \ |K(x)| \leq \min\left\{2M^2, L_2^{-2}\right\}. \tag{70}$$

Furthermore (see [10]), for $r \in (0, L_2)$, the function $\mathrm{d}(x) = \mathrm{d}(x, \partial D)$ is C^2 in $S_r = \overline{D} \setminus (D)_r$ and $|\nabla \mathrm{d}(x)| = 1$ for $x \in S_r$. By simple calculation we have

$$|\partial (D)_r| = |\partial D| + 2r \int\limits_{\partial D} H(x)\, ds + r^2 \int\limits_{\partial D} K(x)\, ds;$$

by the coarea formula we obtain

$$|S_r| = \int\limits_0^r |\{x : \mathrm{d}(x) = t\}|\, dt = r\, |\partial D| + r^2 \int\limits_{\partial D} H(x)\, ds + \frac{r^3}{3} \int\limits_{\partial D} K(x)\, ds.$$

Set $f(r) = |S_r|$. By inequality (see [11])

$$|f'(0)| \leq 12e^2 \|f\|^{\frac{2}{3}} \left(\|f'''\| + \frac{6}{L_2^3} \|f\| \right)^{\frac{1}{3}},$$

where $\|f\| = \max\limits_{[0, L_2]} |f|$, and (70) we obtain

$$|\partial D| \leq 12e^2 |D|^{\frac{2}{3}} \left(4M\, |\partial D| + \frac{6}{L_2^3} |D| \right)^{\frac{1}{3}}; \tag{71}$$

applying the arithmetic–geometric inequality to the right side of (71) we obtain (67).

From

$$|\partial [D]_r| = |\partial D| - 2r \int\limits_{\partial D} H(x)\, ds + r^2 \int\limits_{\partial D} K(x)\, ds$$

and (67) we obtain (68).

Finally, from

$$|[D]_r \setminus D| = r \left(|\partial D| - r \int\limits_{\partial D} H(x)\, ds + \frac{r^2}{3} \int\limits_{\partial D} K(x)\, ds \right)$$

and recalling (70) and (67) we have (69). □

Proposition 3.5. *Let D_1, D_2 satisfy* (7). *The following inequality then holds:*

$$\delta_{\mathcal{H}}\left(D_1, D_2\right) \leq 2dL_1^{-\frac{3}{4}} \left|D_1 \vartriangle D_2\right|^{\frac{1}{4}}. \tag{72}$$

Proof. Denote $\sigma = \left|D_1 \vartriangle D_2\right|$. First suppose that

$$\sigma \leq \frac{\pi L_1^3}{50}. \tag{73}$$

Let $x \in \overline{D}_1 \setminus D_2$. By the interior sphere property of D_1 (see Proposition 2.2), it easily follows that there exists a ball $\mathrm{B}\big(p, \frac{L_1}{2}\big)$ such that $\mathrm{B}\big(p, \frac{L_1}{2}\big) \subset D_1$ and $x \in \partial \mathrm{B}\big(p, \frac{L_1}{2}\big)$. We have

$$\left|B\left(p, \frac{L_1}{2}\right) \setminus D_2\right| \leq \left|D_1 \setminus D_2\right| \leq \sigma.$$

By (73) we have that $\mathrm{B}\big(p, \frac{L_1}{2}\big) \cap D_2$ is nonempty and there exists $y \in \partial D_2 \cap \overline{B\left(p, \frac{L_1}{2}\right)}$ such that

$$|x-y| = \mathrm{d}\left(x, \overline{D}_2 \cap \overline{B\left(p, \frac{L_1}{2}\right)}\right).$$

By the exterior sphere property of D_2, it follows that there exists a ball $\mathrm{B}(q, L_1)$ contained in $\Omega \setminus \overline{D}_2$ and tangent in y to ∂D_2. By simple geometric considerations we have that $x, y \in \overline{B\left(p, \frac{L_1}{2}\right)} \cap \overline{\mathrm{B}\left(q, L_1\right)}$. Furthermore, the inclusion $\overline{B\left(p, \frac{L_1}{2}\right)} \cap \overline{\mathrm{B}\left(q, L_1\right)} \subset \overline{D}_1 \setminus D_2$ gives

$$\left|\mathrm{B}\left(p, \frac{L_1}{2}\right) \cap \mathrm{B}\left(q, L_1\right)\right| \leq \sigma.$$

This estimate and a bound of diameter of $\mathrm{B}\big(p, \frac{L_1}{2}\big) \cap \mathrm{B}(q, L_1)$ by its volume gives

$$|x-y| \leq \text{diameter of } \left(\mathrm{B}\left(p, \frac{L_1}{2}\right) \cap \mathrm{B}\left(q, L_1\right)\right) \leq 2d\left(L_1^{-3}\sigma\right)^{\frac{1}{4}}.$$

Therefore, if $x \in \overline{D}_1 \setminus D_2$, then

$$\mathrm{d}\left(x, D_2\right) \leq 2d\left(L_1^{-3}\sigma\right)^{\frac{1}{4}}.$$

Since the same estimate is valid for $\mathrm{d}(x, D_1)$, if $x \in \overline{D}_2 \setminus D_1$, (72) holds if (73) is fulfilled. Otherwise, if (73) is not fulfilled, (72) is trivial. □

Acknowledgment. I wish to express my gratitude to Prof. Mario Primicerio, who proposed this problem to me.

REFERENCES

[1] G. Alessandrini, *Stable determination of a crack from boundary measurements*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 497–516.

[2] G. Alessandrini, *Stability for the crack determination problem*, in Proc. Lapland Conference on Inverse Problems, Inari, 1992.

[3] S. Andrieux, A. Ben Abda, and M. Jaoua, *Identifiabilitè de frontière inaccessible par des mesures de surface*, C.R. Acad. Sci. Paris, 316 (1993), pp. 429–434.

[4] A. Ben Abda, *Sur Quelques Problemes Inverses Geometriques via des Equations de Conduction Elliptiques: Etude Theorique et Numerique*, Ph.D. thesis, Ecole National d'ingenieurs de Tunis, 1993.

[5] J. R. Cannon, *The one dimensional heat equation*, Addison–Wesley, Reading, MA, 1984.

[6] A. Elayyan and V. Isakov, *On uniqueness of recovery of the discontinuos conductivity coefficient of a parabolic equation*, SIAM J. Math. Anal., 28 (1997), pp. 49–59.

[7] L. C. Evans and R. E. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[8] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[9] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Math. J., 38 (1989), pp. 527–556.

[10] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.

[11] A. Gorny, *Contribution a l'étude des functions derivables d'une variable réelle*, Acta Math., 71 (1939), pp. 317–358.

[12] P. G. Kaup and F. Santosa, *Nondestructive evaluation of corrosion damage using electrostatic measurements*, J. Nondestructive Eval., 1995.

[13] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishatskii, *Ill-posed problems of mathematical physics and analysis*, Transl. Math. Monograph 64, Amer. Math. Soc., Providence, RI, 1986.

[14] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type*, Transl. Math. Monograph 23, Amer. Math. Soc., Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1371–1388, November 1997

GLOBAL CONVEXITY IN A THREE-DIMENSIONAL INVERSE ACOUSTIC PROBLEM[*]

MICHAEL V. KLIBANOV[†]

Abstract. We consider an inverse scattering problem (ISP) for the acoustic equation $u_{tt} = c^2(x)\Delta u, u|_{t=0} = 0, u_t|_{t=0} = \delta(x), x \in \mathbb{R}^3$. The ISP consists of the determination of the speed of sound $c(x)$ inside a bounded domain $\Omega \subset \mathbb{R}^3$ given $c(x)$ outside Ω and measurements of the amplitude $u(x,t)$ of the sound at the boundary $\partial\Omega$, $u|_{\partial\Omega} = \varphi(x,t)$. This problem is nonoverdetermined since only a single source location at $\{0\}$ is counted. Assuming regularity of the rays generated by $c(x)$ and using the *Carleman's weight functions*, we construct a cost functional J_λ. The main result is Theorem 3.1, which claims global strict convexity of J_λ on "reasonable" compact sets of solutions. Therefore, global convergence on such a set of a number of standard minimization algorithms to the unique global minimum of J_λ (i.e., solution of the ISP) is guaranteed. This in turn shows a possibility of constructions of numerical methods for this ISP which would not be affected by the problem of local minima.

Key words. inverse scattering problem, Carleman estimate, globally convex cost functional

AMS subject classifications. 35L15, 65N12, 65M30

PII. S0036141096297364

1. Introduction.

Let $c(x)$ be the speed of sound at the point $x \in \mathbb{R}^3, c(x) \geq$ const > 0. The amplitude $u(x,t)$ of the sound waves propagating in $\mathbb{R}^3$ from the point source located at $\{0\}$ is governed by the wave equation

$$\begin{gathered} u_{tt} = c^2(x) \bigtriangleup u,\ x \in \mathbb{R}^3,\ t \in (0,T), \\ u|_{t=0} = 0,\ u_t|_{t=0} = \delta(x). \end{gathered} \tag{1.1}$$

Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with a piecewise smooth boundary $\partial\Omega$ and $\{0\} \notin \overline{\Omega}$. We work with the following inverse scattering problems (ISPs).

ISP 1. *Assume that function $c(x)$ is given outside of Ω and is unknown inside of Ω. Determine $c(x)$ for $x \in \Omega$ assuming that the following function $\varphi(x,t)$ is given:*

$$u|_{\partial\Omega\times(0,T)} = \varphi(x,t).$$

Function φ can be interpreted as the result of measurements of the sound's amplitude. This problem has well-known applications in geophysics and ocean acoustics.

Solving the outer boundary value problem

$$\begin{aligned} u_{tt} &= c^2(x) \bigtriangleup u,\ \ x \in \mathbb{R}^3\backslash\Omega,\ t \in (0,T), \\ u|_{t=0} &= 0,\ u_t|_{t=0} = \delta(x), \\ u|_{\partial\Omega\times(0,T)} &= \varphi(x,t), \end{aligned}$$

one can uniquely determine function $u(x,t)$ outside of Ω. Hence, the following function $\xi(x,t)$ is given as well:

$$\frac{\partial u}{\partial n}|_{\partial\Omega\times(0,T)} = \xi(x,t).$$

[*] Received by the editors January 17, 1996; accepted for publication (in revised form) June 26, 1996.
http://www.siam.org/journals/sima/28-6/29736.html

[†] Department of Mathematics, University of North Carolina at Charlotte, Charlotte, NC 28223 (mklibanv@unccsun.uncc.edu).

Now, with the case of backscattering data in mind, we consider a more complicated ISP. Namely, let $\omega \subset \partial\Omega$ be a part of $\partial\Omega$.

ISP 2. *Assume that function $c(x)$ is given outside of Ω and is unknown inside of Ω. Determine $c(x)$ for $x \in \Omega$ assuming that the following two functions are given:*

$$u|_{\partial\Omega\times(0,T)} = \varphi(x,t), \ \ \frac{\partial u}{\partial n}|_{\omega\times(0,T)} = \xi(x,t), \ t \in (0,T). \tag{1.2}$$

As follows from our method, precise values of the function φ at $\partial\Omega\backslash\omega = \omega_1$ are not important. Rather, bounds for φ and its derivatives up to the second order are important at ω_1 in order to keep $u(x,t)$ bounded. Therefore, one can consider ISP 2 as an ISP with the *backscattering* data only.

Remarks (also, see a relevant discussion in section 5). (i) In the case of backscattering data, the function φ is usually only given at $\omega\times(0,T)$ rather than at $\partial\Omega\times(0,T)$. However, it follows from our method that one can prescribe reasonable "pseudo" values for this function at $(\partial\Omega\backslash\omega)\times(0,T)$. Given the right choice of parameters λ and ν in the Carleman's weight function (Theorem 3.1), these pseudovalues will provide very little impact on the solution $c(x)$ at the points x located far from $\partial\Omega\backslash\omega$. The real role of these pseudovalues is to provide an upper bound for the solution. In terms of estimates given below, we would not be able to obtain the "Lipshitz-like" Carleman estimate in the form of Theorem 4.1 without a boundary condition at $\partial\Omega\backslash\omega$. Rather, we would obtain a "Hölder-like" Carleman estimate. To see this, one might compare Theorem 8.3.1 in [3] with Theorem 3 in Chapter 4, section 1 in [8]. In [3], the function $u \in C_0^\infty(\Omega)$ leads to the Lipshitz-like Carleman estimate, whereas in [8], u does not have zero boundary values, which leads to a Hölder estimate. The difference between Theorem 4.1 below and Theorem 8.3.1 in [3] is that, unlike [3], we assign a single boundary condition only at $\partial\Omega\backslash\omega$.

(ii) Even in the case of backscattering data, the function $\xi(x,t)$ in (1.2) can often be computed, rather than given a priori. The following example clarifies this statement. Suppose that the function $c(x)$ is given in the half-space $\{x_3 > 0\}$ and is unknown for $\{x_3 < 0\}$. Also, let the function $\varphi(x,t) = u|_{\{x_3=0\}\times(0,T)}$ be given for all $(x,t) \in \{x_3 = 0\}\times(0,T)$. Then, function $\xi(x,t)$ can be computed for $\{x_3 = 0\}\times(0,T)$ as the solution of the corresponding boundary value problem in the half-space. Let R be a positive constant. Assume that one wants to find function $c(x)$ for $x \in \{|x| < R, \ x_3 < 0\}$ only, rather than for all $x \in \{x_3 < 0\}$. Denote

$$\Omega = \{|x| < 2R, \ x_3 < 0\}, \ \omega = \{|x| < 2R, \ x_3 = 0\}, \ \partial\Omega' = \Omega\backslash\omega.$$

Next, assign reasonable "pseudo" values for the function

$$\varphi(x,t)|_{\partial\Omega'\times(0,T)} \ = u|_{\partial\Omega'\times(0,T)}.$$

Then, given the right choice of parameters λ and ν in the Carleman's weight function in (3.2), our method should provide a good approximation for the function $c(x)$ for $x \in \{|x| < R, \ x_3 < 0\}$.

A variety of numerical methods for different versions of multidimensional coefficient ISPs including inverse acoustic problems has been developed in the past; cf. [1] and references cited therein. However, in nonlinear situations, these methods do not guarantee the absence of the local minima of cost functionals. In this paper we construct such a cost functional J_λ for the ISP 2, which is globally strictly convex on "reasonable" compact sets of solutions.

Thus, J_λ does not have local minima in the interior of such a compact set. Further, by Tikhonov's principle [9], one should assume that a solution of the ISP 2 exists and belongs to the interior of a given compact set, at least in the case of noiseless data. Hence J_λ attains its unique global minimum at an interior point of this compact set. Finally, in the case of a presence of a small-level noise in the data, convexity of J_λ implies that its unique global minimum is attained at a point which is close to the solution for the noiseless data (the corresponding result easily can be derived using the general framework of the theory of ill-posed problems [1, 9]).

Our main idea consists of the use of the *Carleman's weight functions* $\alpha^2(x)$, which are involved in the Carleman estimates for Laplace's operator (see Theorem 4.1). For this reason, we call our technique *Carleman's weight method* (CWM). The major "price" for our method is an assumption that a function associated with $u(x,t)$ has a finite number N of Fourier harmonics with respect to t for an orthonormal basis in L_2-space. That is, we work with the Galerkin approximation without proof of its convergence as $N \to \infty$. Thus, N is the regularization parameter for this ill-posed problem. Note that such an assumption is usually quite acceptable in numerical methods. In fact, all the preceding algorithms deal with some similar assumptions (implicitly or explicitly). We also note that the proof of convergence of our technique as $N \to \infty$ would almost inevitably lead to the proof of a global uniqueness theorem for this ISP. The latter, however, is a long-standing unsolved problem; cf. [7, 8].

The global convexity of cost functionals for similar 3-dimensional hyperbolic and parabolic ISPs was established by CWM in [2, 5]. However, only coefficients at the lower order derivatives were reconstructed in these references.

CWM consists of two main steps. First, by eliminating the unknown coefficient, one obtains a special boundary value problem for an elliptic system of nonlinear PDEs. Second, to solve this system, one constructs a cost functional J_λ. The presence of the Carleman's weight function in J_λ together with the Carleman estimate for Laplace's operator ensures the global strict convexity of J_λ.

In section 2, we obtain a nonlinear elliptic system convenient for treatment by CWM. In section 3, we construct the cost functional J_λ and prove its convexity. In section 4, we derive the Carleman estimate. Section 5 is devoted to discussion.

2. Nonlinear elliptic system convenient for CWM.

In this section, we use the detailed study of propagation of singularities of the hyperbolic Cauchy problem (1.1) being undertaken in the book [8]. For $x_0, x \in \mathbb{R}^3$, let $\tau(x, x_0)$ be the travel time of the sound traveling from x_0 to x. Then, function $\tau(x, x_0)$ satisfies the eikonal equation

$$|\bigtriangledown_x \tau|^2 = \frac{1}{c^2(x)}. \tag{2.1}$$

This equation generates the family of rays $L(x, x_0)$ along which the first arrival signal travels from x_0 to x. One also calls these rays *geodesic lines* generated by Riemann's matrix

$$d\tau = \frac{1}{c(x)}\sqrt{(dx_1)^2 + (dx_2)^2 + (dx_3)^2}.$$

The following functional $\tau(M)$ attains its minimum value on the geodesic line $L(x, x_0)$ (this is Fermat's principle):

$$\tau(M) = \int\limits_{M(x,x_0)} \frac{ds}{c(x)},$$

where $M(x, x_0)$ is a smooth curve connecting x_0 with x.

Regularity assumption. *In what follows, we will always assume that the family of geodesic lines is regular in $\mathbb{R}^3$; i.e., for every two points x_0, $x \in \mathbb{R}^3$, there exists a unique geodesic line $L(x, x_0)$ connecting them.*

Denote $\tau_0(x) = \tau(x, 0)$. The surface $\{t = \tau_0(x)\}$ in $\mathbb{R}^3 \times (0, T)$ defines the characteristic cone for the solution of the hyperbolic Cauchy problem (1.1). So, $u(x, t) = 0$ for $t < \tau_0(x)$. Thus, we will consider function $u(x, t)$ above this cone only. Let $G = \{(x, t) : x \in \Omega, \ \tau_0(x) \leq t \leq T\}$. Function $u(x, t)$ consists of singular and regular parts which we denote u_0 and u_1, respectively, where $u = u_0 + u_1$. CWM requires that

$$u_1(x, t) \in C^2(\overline{G}), \ \tau_0(x) \in C^2(\overline{\Omega}). \tag{2.2}$$

By [8] and to ensure (2.2), we impose somewhat excessive smoothness conditions on the function $c(x)$ (these conditions might likely be relaxed in practical computations). Theorem 4.1 of [8] can be reformulated as follows in our case.

Theorem 2.1. *Let $c(x) \geq \text{const} > 0$, $0 \notin \overline{\Omega}$, $c(x) \in C^{l+4}(\mathbb{R}^3)$, $l \geq 2s + 7$ (where $s \geq -1$ is an integer), and $H(t)$ be the Heaviside function. Then the solution of the Cauchy problem* (1.1) *has the form*

$$u(x, t) = \frac{\sigma_{-1}(x)}{\tau_0(x)} \delta(t - \tau_0(x)) + H(t) \sum_{k=0}^{s} \sigma_k(x) H_k(t^2 - \tau_0^2(x)) + v_s(x, t),$$

where

$$H_k(t) = \frac{t^k}{k!} H(t), \ k \geq 0, \ \sigma_k(x) \in C^{l-2k}(\overline{\Omega}) \text{ for } k = -1, ..., s,$$
$$\tau_0 \in C^{l+3}(\overline{\Omega}), \textit{ and } \sigma_{-1}(x) > 0 \textit{ for all } x.$$

Furthermore, $v_s(x, \tau_0(x)) = 0$; for $s \geq 1$, function $v_s(x, t)$ is continuous in $\overline{G}$ with its derivatives $\mathcal{D}_\chi^\alpha \mathcal{D}_t^\beta \mathcal{V}_s$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $|\alpha| + \beta \leq s - 1$.

Hence, the singular part of the function u is

$$u_0(x, t) = \frac{\sigma_{-1}(x)}{\tau_0(x)} \delta(t - \tau_0(x)), \ \ \sigma_{-1}(x) > 0, \tag{2.3}$$

and the regular part is

$$u_1(x, t) = H(t) \sum_{k=0}^{s} \sigma_k(x) H_k(t^2 - \tau_0^2(x)) + v_s(x, t).$$

Thus, to ensure that $u_1 \in C^2(\overline{G})$, we take $s = 3, \ l = 13,$ and

$$c \in C^{17}(\mathbb{R}^3). \tag{2.4}$$

Remark. Because of the singular part (2.3) in the function $u(x, t)$, functions $\tau_0(x)|_{\partial\Omega}$ and $\frac{\partial \tau_0(x)}{\partial n}|_\omega$ easily can be determined from $u(x, t)|_{\partial\Omega \times (0,T)}$ and $\frac{\partial u}{\partial n}(x, t)|_{\omega \times (0,T)}$.

Now we replace the cone $\{t = \tau_0(x)\}$ with the hyperplane $\{t = 0\}$. To do this, we introduce a new function $W(x, t)$:

$$W(x, t) = \int_0^t dz \int_0^z u(x, y + \tau_0(x)) dy. \tag{2.5}$$

Here the integral of the singular part u_0 of the function u is understood in the conventional setting as (see (2.3))

$$\int_0^z u_0(x, y + \tau_0(x))dy = \frac{\sigma_{-1}(x)}{\tau_0(x)} H(z).$$

Hence,

$$\lim_{t\to 0^+} W(x,t) = 0, \ \lim_{t\to 0^+} \frac{\partial W}{\partial t}(x,t) = \frac{\sigma_{-1}(x)}{\tau_0(x)} > 0. \tag{2.6}$$

Further, let $T_1 = T - \max\limits_{\Omega} \tau_0(x)$. We assume that $T_1 > 0$. Let $\Omega_{T_1} = \Omega \times (0, T_1)$; then

$$W \in C^2(\overline{\Omega}_{T_1}). \tag{2.7}$$

Because of (1.1), (1.2), and (1.5), function W satisfies the following conditions:

$$\triangle W - 2 \bigtriangledown (W_t) \bigtriangledown \tau_0 - W_t \triangle \tau_0 = 0 \text{ in } \Omega_{T_1}, \tag{2.8a}$$

$$W(x,0) = 0, \tag{2.8b}$$

$$W_t(x,0) > 0 \text{ in } \overline{\Omega}, \tag{2.8c}$$

$$W|_{\partial\Omega\times(0,T_1)} = \varphi_1(x,t), \ \frac{\partial W}{\partial n}|_{\omega\times(0,T_1)} = \xi_1(x,t), \tag{2.8d}$$

where functions φ_1 and ξ_1 are generated by the functions φ and ξ due to the transformation (2.5).

An inconvenience of equation (2.8a) is that highest order derivatives fall on two unknown functions rather than just a single one: $W(x,t)$ and $\tau_0(x)$. So, to obtain a *system* of the equations with respect to τ_0 and W, we take $t = 0$ in (2.8a). Then, using (2.8b) and (2.8c), we obtain

$$\triangle \tau_0 + \frac{2}{W_t(x,0)} \bigtriangledown [W_t(x,0)] \bigtriangledown \tau_0 = 0. \tag{2.9}$$

To find the Dirichlet and Neumann data for $\tau_0(x)$, we analyze singularities of functions φ and ξ in (1.2) at $t \to \tau_0^+(x)$ due to the propagation of the singularity in (2.3). Hence, the following two functions are known as well:

$$\tau_0 \mid_{\partial\Omega} = p(x), \ \frac{\partial \tau_0}{\partial n}|_{\omega} = q(x). \tag{2.10}$$

We cannot solve the nonlinear system (2.8)–(2.10) in its present form. So, we simplify this system by cutting off Fourier harmonics for $W(x,t)$ with respect to t. Let $\{a_n(t)\}_{n=1}^{\infty} \subset C^2[0, T_1]$ be an orthonormal basis in $L_2[0, T_1]$ such that $a_n(0) = 0$. Let $\{Q_n(x)\}_{n=1}^{\infty}$ be Fourier coefficients of the function $W(x,t)$ with respect to this basis,

$$Q_n(x) = \int\limits_0^{T_1} W(x,t) a_n(t) dt.$$

Choose an integer $N > 1$; we then impose the following assumption.

MAIN ASSUMPTION. *Denote*

$$v(x,t) = \sum_{n=1}^{N} a_n\ (t)\ Q_n\ (x),$$
$$\widetilde{\varphi}_1(x,t) = \sum_{n=1}^{N} a_n\ (t)\ \varphi_{1n}\ (x),$$
$$\widetilde{\xi}_1(x,t) = \sum_{n=1}^{N} a_n\ (t)\xi_{1n}(x),$$

where $\varphi_{1n}(x)$ *and* $\xi_{1n}(x)$ *are Fourier coefficients of the corresponding functions in* (2.8d)*. So, we assume that function* v *satisfies exactly the same conditions* (2.8a)–(2.8c)*,* (2.9)*, and* (2.10) *as does function* W*, whereas* (2.8d) *is replaced with*

$$v|_{\partial\Omega\times(0,T_1)} = \widetilde{\varphi}_1(x,t),\ \frac{\partial v}{\partial n}|_{\omega\times(0,T_1)} = \widetilde{\xi}_1(x,t).$$

Therefore, we arrive at the following nonlinear system of PDEs:

$$\triangle v - 2 \bigtriangledown (v_t) \bigtriangledown \tau_0 - 2v_t \triangle \tau_0 = 0, \tag{2.11a}$$

$$v(x,0) = 0, \tag{2.11b}$$

$$v_t(x,0) > 0, \tag{2.11c}$$

$$\triangle\tau_0 + \frac{2}{v_t(x,0)} \bigtriangledown [v_t(x,0)] \bigtriangledown \tau_\partial = 0, \tag{2.11d}$$

$$v|_{\partial\Omega} = \widetilde{\varphi}_1,\ \frac{\partial v}{\partial n}|_\omega = \widetilde{\xi}_1,\ \tau_0|_{\partial\Omega} = p,\ \frac{\partial \tau_0}{\partial n}|_\omega = q. \tag{2.11e}$$

Remark. Because of the finite-dimensional Fourier approximation, the existence of a solution to the system (2.8a)–(2.8d) does not necessarily imply the existence of a solution to the system (2.11a)–(2.11e). This is a major shortcoming of our main assumption, since we cannot evaluate our method as $N \to \infty$ (also, see the discussion below). However, we think that the finite-dimensional time dependence assumption is acceptable for numerical methods. In addition, Theorem 3.1 guarantees that our method finds a vector-valued function (v, τ_0), which provides a minimal discrepancy between left- and right-hand sides of (2.11a) and (2.11d), provided, of course, that a solution of the minimization problem (3.2) does exist and belongs to the interior of the set K_2 (this last assumption is very similar to the classical Tikhonov's "compact set principle"; see also the remark after the statement of Theorem 3.1).

Further, introduce vectors

$$A = \left(a_1^{'}(0), ..., a_N^{'}(0)\right),\ Q(x) = (Q_1(x), ..., Q_N(x)),$$
$$\Phi(x) = (\varphi_{11}(x), ..., \varphi_{1N}(x)),\ \text{and}\ \Psi(x) = (\xi_{11}(x), ..., \xi_{1N}(x)).$$

Let $[\ ,\ ]$ denote the dot product in $\mathbb{R}^N$. Choose an integer $n,\ 1 \leq n \leq N$, multiply both sides of (2.11a) by $a_n(t)$, and integrate it over the time interval $(0, T_1)$. We obtain the following nonlinear elliptic system of the second order:

$$\triangle Q - 2 \bigtriangledown (BQ) \cdot \bigtriangledown \tau_0 + 2BQ \frac{\bigtriangledown [A, Q] \cdot \bigtriangledown \tau_0}{[A, Q]} = 0,$$

$$\triangle \tau_0 + 2 \frac{\bigtriangledown [A, Q] \cdot \bigtriangledown \tau_0}{[A, Q]} = 0, \tag{2.12}$$

$$Q|_{\partial\Omega} = \Phi(x),\ \ \frac{\partial Q}{\partial n}|_{\omega} = \Psi(x),\ \tau_0|_{\partial\Omega} = p(x),\ \ \frac{\partial \tau_0}{\partial n}|_{\omega} = q(x),$$

where B is an $N \times N$ matrix with the elements

$$b_{nk} = \int_0^{T_1} a_n(t) a_k^{'}(t) dt;\ n, k = 1, ..., N.$$

While the form of the system (2.12) is almost what we need, it is still a bit inconvenient because of the nonzero boundary conditions. Hence, to obtain zero boundary conditions, we assume that there exists a given N-dimensional vector-valued function $F(x)$ and a given real-valued function $g(x)$ such that

$$F = [F_1, (x), ..., F_N(x)],\ F_i, g \in C^2(\overline{\Omega}),$$

$$F|_{\partial\Omega} = \Phi(x),\ \frac{\partial F}{\partial n}|_{\omega} = \Psi(x), \tag{2.13}$$

$$g|_{\partial\Omega} = p(x),\ \frac{\partial g}{\partial n}|_{\omega} = q(x).$$

Denote

$$P(x) = Q(x) - F(x),\ r(x) = \tau_0(x) - g(x).$$

Then (2.12) and (2.13) lead to

$$\triangle P - 2 \bigtriangledown (BP) \bigtriangledown (r + g) - 2 \bigtriangledown (BF) \bigtriangledown r$$

$$+2\ B(P + F)\ \frac{\bigtriangledown [A, P + F] \bigtriangledown (r + g)}{[A, P + F]} = R^{(1)}(x),$$

$$\triangle r + 2 \frac{\bigtriangledown [A, P + F] \cdot \bigtriangledown (r + g)}{[A, P + F]} = R^{(2)}(x), \tag{2.14}$$

$$P|_{\partial\Omega} = \frac{\partial P}{\partial n}|_{\omega} = 0,\ \ r|_{\partial\Omega} = \frac{\partial r}{\partial n}|_{\omega} = 0,$$

where

$$R^{(1)}(x) = -\left(\triangle F - 2 \bigtriangledown (BF) \bigtriangledown g\right),\ R^{(2)}(x) = -\triangle g. \tag{2.15}$$

So, our goal below is to solve the system (2.14), (2.15). Given the solution (P, r) of this system, one easily can reconstruct the function $\tau_0(x)$ and, therefore, $c(x)$. We also note that while zero boundary conditions for the vector (P, r) are convenient for theoretical analysis, in practical computations, one can likely work with the nonzero conditions for the vector (Q, τ_0) in (2.12).

3. Uniformly strictly convex cost functional. To solve the system (2.14) and (2.15), we introduce a weighted cost functional J_λ. As the weight function, we choose the *Carleman's weight function*, i.e., one involved in the Carleman estimate for the Laplacian operator (see Theorem 4.1).

Choose a point $x_0 \in \mathbb{R}^3 \backslash \overline{\Omega}$ such that there exists a part $\omega_1 \subseteq \omega$ for which

$$(x - x_0, n(x)) < 0 \text{ for all } x \in \omega_1$$

and

$$(x - x_0, n(x)) \geq 0 \text{ for all } x \in \partial\Omega \backslash \omega_1,$$

where $n(x)$ is the outward pointing unit normal to $\partial\Omega$ and $(\ ,\ )$ is the dot product in $\mathbb{R}^3$. Without loss of generality, we will always assume that $\Omega \subset \left\{|x - x_0| < \frac{1}{2}\right\}$. Consider functions

$$\psi(x) = |x - x_0|^2 + \frac{1}{4}, \ \ \alpha(x) = \exp\left[\lambda \psi^{-\nu}\right],$$

where λ and ν are large positive parameters to be chosen later. Because of the Carleman estimate of Theorem 4.1, we call $\alpha(x)$ *Carleman's weight function.* Let x_1 be the point on ω_1 closest to x_0. It is important for our method that Carleman's weight $\alpha(x)$ attains its maximum value at x_1 and decays exponentially in Ω with respect to the distance from x_1. The level surfaces of $\alpha(x)$ are spheres with the center at x_0.

Remark. It can be shown that if, for example, Ω is a cube, $\Omega = (1, a)^3$ with $a = \text{const} > 1$ and $\omega = \{x_3 = 1\} \cap \partial\Omega$, then one can choose $\alpha(x)$ in a simpler way as $\alpha(x) = \exp(\lambda x_3^{-\nu})$ with its level surfaces $\{x_3 = \text{const}\}$.

Let $H_0^2(\Omega)$ be the subspace of $H^2(\Omega)$ consisting of all the real-valued functions $f(x)$ satisfying the boundary conditions

$$f|_{\partial\Omega} = \left.\frac{\partial f}{\partial n}\right|_{\omega} = 0.$$

We will say that a k-dimensional vector-valued function $\beta(x)$ belongs to $H_0^2(\Omega)$ if all of its components belong to $H_0^2(\Omega)$. The same is true for any other Banach space which will be used below. For the norm $\| \cdot \|$ of such a Banach space, we define the norm of β as

$$\| \beta \| = \left[\sum_{i=1}^{k} \| \beta_i \|^2\right]^{\frac{1}{2}}.$$

Because of (2.14), denote

$$R(x) = \left(R^{(1)}(x), R^{(2)}(x)\right), \ S(x) = (P(x), r(x)),$$

$$\begin{aligned} V_1(S) = \triangle P - 2 \bigtriangledown (BP) \bigtriangledown (r + g) - 2 \bigtriangledown (BF) \bigtriangledown r \\ + 2B(P + F) \frac{\bigtriangledown [A, P + F] \bigtriangledown (r + g)}{[A, P + F]}, \end{aligned} \tag{3.1}$$

$$V_2(S) = \triangle r + 2\frac{\nabla [A, P + F] \nabla (r + g)}{[A, P + F]}, \text{ and } V(S) = (V_1(S), V_2(S)).$$

Clearly, $S \in H_0^2(\Omega)$ and $V(S)$ is an $(N+1)$-dimensional vector. Introduce the cost functional $J_\lambda(S)$ as

$$J_\lambda(S) = \int_\Omega [V(S) - R]^2 \alpha^2 dx, \tag{3.2}$$

where

$$[V(S) - R]^2 = [V(S) - R, V(S) - R]$$

is the square length of the $(N+1)$-dimensional vector $V(S) - R$. So, by (2.14) and (2.15), one should find such an $(N+1)$-dimensional vector-valued function $S = \widetilde{S} \in H_0^2(\Omega)$ that provides the minimum of the functional J_λ,

$$\min_S J_\lambda(S) = J_\lambda(\widetilde{S}).$$

First, suppose that the data F and g are given without noise and that $S^* \in H_0^2(\Omega)$ is a solution of the equation $V(S) - R = 0$. Then,

$$\min_S J_\lambda(S) = J_\lambda(S^*) = 0.$$

It also follows from the strict convexity of J_λ (Theorem 3.1) that even if functions F and g are given with noise, which is sufficiently small in the $H^2(\Omega)$-norm, then the point $\widetilde{S}$ of the unique global minimum of J_λ (on the compact set K_2 introduced below) is close to the solution S^* for the noiseless data, provided, of course, that S^* exists. The existence of S^*, however, should be assumed a priori by Tikhonov's approach to ill-posed problems; cf. [1, 9].

By Tikhonov's principle [9], we will minimize J_λ over "reasonable" compact sets introduced below.

In what follows, C will denote different positive constants depending only on Ω, ω_1, and x_0. Let $m_1 = \text{const} > 0$ and $K_1 = K_1(m_1) \subset C^2(\overline{\Omega})$ be the set of vector-valued functions (F, g) such that

$$\| (F, g) \|_{C^2(\overline{\Omega})} \leq m_1 . \tag{3.3}$$

Hence,

$$\| R \|_{C(\overline{\Omega})} \leq C m_1. \tag{3.4}$$

Likewise, we want to bound from the above C^1-norms of all vector-valued functions $S = (P, r)$ under consideration. Therefore, let m_2 and m_3 be two positive constants, $m_2 < m_3$. Then $K_2 = K_2(m_2, m_3)$ will denote a compact set of vector-valued functions $S = (P, r) \in H_0^2(\Omega) \cap H^3(\Omega)$ such that

$$[A, P + F] \geq m_2 \text{ for all } (F, g) \in K_1 \text{ and all } (P, r) \in K_2 \tag{3.5}$$

and

$$\| S \|_{H^3(\Omega)} \leq m_3. \tag{3.6}$$

Note that because of (2.11c), $[A, P+F] > 0$, which explains (3.5). Inequalities (3.5) and (3.6) imply that K_2 is a *convex* set in $H_0^2(\Omega)$. In addition, by Sobolev's embedding theorem, $K_2 \subset C^1(\overline{\Omega})$ and

$$\| S \|_{C^1(\overline{\Omega})} \leq CNm_3 \text{ for all } S \in K_2. \tag{3.7}$$

So, in working on a computational implementation of our method, one should assume, by Tikhonov's principle, that in the case of noiseless data, solution S^* of problem (2.14) does exist and belongs to the interior of K_2 [9].

The following theorem is the main result of this paper.

THEOREM 3.1. *Let Ω be a convex bounded domain, $0 \notin \overline{\Omega}$, the speed of sound $c(x) \geq \text{const} > 0$, $c(x) = \text{const}$ outside of Ω, $c(x) \in C^{17}(\mathbb{R}^3)$, the family of geodesic lines generated by $c(x)$ be regular in $\mathbb{R}^3$, and $\tau_0(x) = \tau(x,0)$ be the travel time from the source $\{0\}$ to the point x. Then there exist large positive parameters $\nu_0 = \nu_0(\Omega,\omega)$ and $\lambda_0 = \lambda_0(\Omega,\omega,N,K_1,K_2)$ such that for $\nu = \nu_0$, all $\lambda \geq \lambda_0$, and for all $(F,g) \in K_1$, the cost functional $J_\lambda(S)$ in (3.2) is uniformly strictly convex on the convex compact set $K_2 \subset H_0^2(\Omega)$. That is, for all $(F,g) \in K_1$ and $S,(S+h) \in K_2$, where $||h||_{H^2(\Omega)} < 1$, $\nu = \nu_0$, and $\lambda \geq \lambda_0$, the following inequality is valid:*

$$J_\lambda(S+h) - J_\lambda(S) - J'_\lambda(S)(h) \geq C_1 \| h \|^2_{H^2(\Omega)},$$

where J'_λ is the Frechet derivative of J_λ at the point S and C_1 is a positive constant, $C_1 = C_1(\Omega,\omega,N,\nu_0,\lambda,K_1,K_2)$, but does not depend on$S,h,F,$ and g.

Remark. There is a question, of course, on how "big" the set K_2 is. First, K_2 is a compact set in $H^2(\Omega)$, which suits the Tikhonov's principle [9] well. Second, the "size" of K_2 depends on the constants m_2 and m_3 in (3.5), (3.6). So, given m_2, the size of K_2 can be big if m_3 is a big number. The bigger the size of K_2, the larger the value of λ_0 one should use in Theorem 3.1.

Proof. In this proof, C_2 will denote different positive constants depending on the sets K_1 and K_2 only, i.e., on the constants $m_1, m_2,$ and m_3 in (3.3)–(3.6). But constants C_2 do not depend on $\lambda, \nu, S,$ and h.

The proof consists of two steps. First, using algebraic manipulations and the Cauchy–Schwarz inequality, we prove that

$$J_\lambda(S+h) - J_\lambda(S) - J'_\lambda(S)(h) \geq \frac{1}{2}\int\limits_\Omega [\triangle h]^2 \alpha^2 dx - C_2 \int\limits_\Omega \left(| \bigtriangledown h|^2 + [h]^2\right)\alpha^2 dx. \tag{3.8}$$

Second, the Carleman estimate of Theorem 4.1 implies that the first term of the right-hand side of (3.7) dominates the rest.

To establish (3.8), we evaluate $J_\lambda(S+h) - J_\lambda(S)$ and single out the derivative $J'_\lambda(S)(h)$. By (3.2),

$$J_\lambda(S+h) - J_\lambda(S) = \int\limits_\Omega [V(S+h) - V(S),\ V(S+h) + V(S) - 2R]\alpha^2 dx.$$

Let $h = (h_1, h_2)$, where h_1 and h_2 represent P and r, respectively. Then a routine algebraic analysis of (3.1) implies that

$$V_1(S+h) - V_1(S) = L_1(S,h) + G_1(S,h),$$

where L_1 and G_1 are the linear and nonlinear parts, respectively, of this difference with respect to h. Namely,

$$L_1(S,h) = \triangle h_1 - 2 \bigtriangledown (Bh_1) \bigtriangledown (r+g) - 2 \bigtriangledown (BP) \bigtriangledown h_2$$

$$+2(Bh_1)\ \frac{\bigtriangledown[A,P+F]\bigtriangledown(r+g)}{[A,P+F]}$$

$$+2B(P+F)\left\{\frac{\bigtriangledown[A,h_1]\bigtriangledown(r+g)+\bigtriangledown[A,P+F]\bigtriangledown h_2}{[A,P+F]}\right\} \tag{3.9}$$

$$-2B(P+F)\frac{\bigtriangledown[A,P+F]\bigtriangledown(r+g)}{[A,P+F]^2}[A,h_1].$$

Also, G_1 satisfies the following estimate:

$$[G_1(S,h)]\le C_2(|\bigtriangledown h|^2+[h]^2). \tag{3.10}$$

Similarly,

$$V_2(S+h)-V_2(S)=L_2(S,h)+G_2(S,h),$$

where L_2 and G_2 represent the linear and nonlinear parts, respectively, of this difference with respect to h:

$$L_2(S,h)=\triangle h_2+2\ \frac{\bigtriangledown[A,h_1]\bigtriangledown(r+g)+\bigtriangledown[A,P+F]\bigtriangledown h_2}{[A,P+F]}$$

$$-2[A,h_1]\ \frac{\bigtriangledown[A,P+F]\bigtriangledown[r+g]}{[A,P+F]^2}, \tag{3.11}$$

and

$$|G_2(S,h)|\le C_2(|\bigtriangledown h|^2+|h|^2). \tag{3.12}$$

Similarly,

$$V(S+h)+V(S)-2R=2(V(S)-R)+\triangle h+G_3(S,h), \tag{3.13}$$

where

$$[G_3(S,h)]\le C_2(|\bigtriangledown h|^2+|h|^2). \tag{3.14}$$

Again, we remark that constants C_2 are independent on S,h because of (3.7). Let $L(S,h)$ and $G(S,h)$ be $(N+1)$-dimensional vector-valued functions of the form

$$L(S,h)=(L_1(S,h),\ L_2(S,h)),\ \ G(S,h)=(G_1(S,h),\ G_2(S,h)).$$

Hence,

$$L(S,h)=\triangle h+\widetilde{L}(S,h), \tag{3.15}$$

where the linear (with respect to h) operator $\widetilde{L}$ contains only the lower order derivatives of h.

By (3.9)–(3.14),

$$J_\lambda(S+h) - J_\lambda\,(S) = \int\limits_\Omega \left[L(S,h) + G(S,h),\ 2(V(S)-R) + \triangle h + G_3(S,h)\right]\alpha^2 dx.$$

Hence,

$$J_\lambda'(S)(h) = 2\int\limits_\Omega \left[L(S,h),\ V(S)-R\right]\alpha^2 dx$$

and

$$J_\lambda(S+h) - J_\lambda(S) - J_\lambda'(S)(h) = \int\limits_\Omega \left[L(S,h), \triangle h + G_3(S,h)\right]\alpha^2 dx$$

$$+\int\limits_\Omega \left[G(S,h), 2(V(S)-R) + \triangle h + G_3(S,h)\right]\alpha^2 dx.$$

This equation together with (3.15) leads to

$$J_\lambda(S+h) - J_\lambda(S) - J_\lambda^{'}(S)(h) = \int\limits_\Omega [\triangle h]^2\alpha^2 dx + \int\limits_\Omega [\widetilde{L}(S,h), \triangle h + G_3(S,h)]\alpha^2 dx$$

$$+\int\limits_\Omega [\triangle h, G_3]\alpha^2 dx + \int\limits_\Omega [G(S,h), 2(V(S)-R) + \triangle h + G_3(S,h)]\alpha^2 dx. \tag{3.16}$$

Thus, the Cauchy–Schwarz inequality, (3.7), (3.10), (3.12), (3.14), and (3.16) immediately lead to (3.8). Further, (3.8) and the Carleman estimate (Theorem 4.1) imply that

$$J_\lambda(S+h) - J_\lambda(S) - J_\lambda'(S)(h) \geq \frac{C}{2\lambda}\sum_{i,i=1}^{3}\int\limits_\Omega \left[h_{x_ix_j}\right]^2\alpha^2 dx$$

$$+\frac{1}{2}C\lambda\int\limits_\Omega |\bigtriangledown h|^2\alpha^2 dx + \frac{1}{2}C\lambda^3\int\limits_\Omega [h]^2\alpha^2 dx - C_2\int\limits_\Omega \left(|\bigtriangledown h|^2 + [h]^2\right)\alpha^2 dx.$$

Hence, for sufficiently large $\lambda \geq \lambda_0 = \lambda_0\left(\Omega, \omega, N, K_1, K_2\right)$, we obtain

$$J_\lambda(S+h) - J_\lambda(S) - J_\lambda'(S)(h) \geq \frac{C}{2\lambda}\sum_{i,i=1}^{3}\int\limits_\Omega \left[h_{x_ix_j}\right]^2\alpha^2 dx$$

$$+C_2\lambda\int\limits_\Omega [\bigtriangledown h]^2\alpha^2 dx + C_2\lambda^3\int\limits_\Omega [h]^2\alpha^2 dx.$$

Let $\sigma = \max\limits_{\overline{\Omega}} \psi(x)$. Then, the last inequality leads to

$$J_\lambda(S+h) - J_\lambda(S) - J_\lambda'(S)(h) \geq \frac{C_2}{\lambda}\exp[2\lambda\sigma^{-\nu}] \parallel h \parallel_{H^2(\Omega)}^2 = C_1 \parallel h \parallel_{H^2(\Omega)}^2 . \qquad \square$$

Remark. A minor concern can be raised about how to keep H^3-norms of function S bounded (since $S \in K_2 \subset H^3(\Omega)$) while only the H^2-norm of h is involved in this theorem. One can argue, however, that in practical computations one could work with the finite-dimensional approximations of S, where all norms are equivalent in the finite-dimensional spaces.

4. Carleman estimate.

One can find the conventional Carleman estimates in the books [3, 4, 7]. In particular, these estimates were obtained in [3, 4] for very general differential operators, including elliptic ones. The main difference between our case and the conventional one is that we integrate over the *whole* domain Ω, Dirichlet data are given on the whole boundary $\partial\Omega$, and Neumann data are given on its *part* ω only, whereas in the conventional case, either the integration is carried out over a *part* of Ω adjacent with ω (and both Dirichlet and Neumann data are given at ω only as in [7]) or the integration is carried out over the whole domain Ω (and *both* Dirichlet and Neumann data are given at the *whole* boundary $\partial\Omega$, as in [3, 4]). This difference makes it necessary for us to carefully evaluate boundary terms arising after applying the Gaussian formula. Therefore, the major difference between our case and the conventional one lies in the method of evaluation of the boundary terms.

One of the most convenient methods of the derivation of the Carleman estimates which allows one to deal with the nonzero boundary terms is the method of [7]. So, our proof essentially follows [7]. The only two differences are in the analysis of the nonzero boundary conditions and second-order derivatives.

THEOREM 4.1. *There exist sufficiently large positive numbers λ_0, ν_0 depending only on Ω, w_1, and x_0 (see the beginning of section 3 about x_0) such that the following Carleman estimate is valid for $\nu = \nu_0$, for all $\lambda \geq \lambda_0$, and for all functions $u \in H_0^2(\Omega)$:*

$$\int\limits_{\Omega} (\triangle u)^2 \alpha^2 dx \geq \frac{C}{\lambda} \int\limits_{\Omega} \sum_{i,j=1}^{3} \left(u_{x_i x_j}\right)^2 \alpha^2 dx$$
$$+C\lambda \int\limits_{\Omega} | \bigtriangledown u|^2 \alpha^2 dx + C\lambda^3 \int\limits_{\Omega} u^2 \alpha^2 dx,$$

where the positive constant C depends on Ω, ω_1, and x_0 only.

Remark. By a slight modification of the proof of this result, we show that one can choose any $\nu \geq \nu_0$, rather that just fix $\nu = \nu_0$. This leads to an obvious slight change of Theorem 3.1.

Proof. Without loss of generality, we assume that $x_0 = 0$. Let $v(x) = u(x)\alpha(x)$; then $u = v \exp[-\lambda\psi^{-\nu}]$. Hence,

$$\begin{aligned} u_{x_i} &= (v_{x_i} + 2\lambda\nu x_i \psi^{-\nu-1} v) \exp[-\lambda\psi^{-\nu}], \\ u_{x_i x_i} &= \left\{v_{x_i x_i} + 4\lambda\nu x_i \psi^{-\nu-1} v_{x_i} \right. \\ &\quad \left. + \left(2\lambda^2\nu^2 x_i^2 \psi^{-2v-2} + \lambda\nu\psi^{-\nu-1} - \lambda\nu(\nu+1) x_i^2 \psi^{-\nu-2}\right) v\right\} \exp[-\lambda\psi^{-\nu}]. \end{aligned}$$

Hence,

$$(4.1)\ (\triangle u)^2 \psi^{\nu+1} \alpha^2 = \left[\triangle v + 4\lambda\nu\psi^{-\nu-1} \sum_{i=1}^{3} x_i v_{x_i} + 4\lambda^2\nu^2\psi^{-2\nu-2} |x|^2 (1+\gamma) v\right]^2 .$$

In what follows, $\gamma = \gamma(x, \lambda, \nu)$ will denote different $C^1(\overline{\Omega})$-functions such that

$$\lim_{\lambda,\nu\to\infty} \| \gamma(x,\lambda,\nu) \|_{C^1(\overline{\Omega})} = 0.$$

Let

$$z_1 = \triangle v,\ z_2 = 4\lambda\nu\psi^{-\nu-1}\sum_{i=1}^{3} x_i v_{x_i},$$

and

$$z_3 = 4\lambda^2\nu^2\psi^{-2\nu-2}|x|^2(1+\gamma)v. \tag{4.2}$$

Then (4.1) leads to

$$(\triangle u)^2\psi^{\nu+1}\alpha^2 = [(z_1+z_3)+z_2]^2\psi^{\nu+1} \geq 2z_2(z_1+z_3)\psi^{\nu+1}. \tag{4.3}$$

First, we estimate $2z_1z_2\psi^{\nu+1}$. Note that

$$8\lambda\nu x_i v_{x_i} \triangle v = 8\lambda\nu\sum_{j=1}^{3} v_{x_jx_j}v_{x_i}x_i$$

$$=\sum_{j=1}^{3}\frac{\partial}{\partial x_j}\left(8\lambda\nu x_i v_{x_j}v_{x_i}\right) - \sum_{j=1}^{3} 8\lambda\nu v_{x_j}v_{x_ix_j}x_i - 8\lambda\nu v_{x_i}^2$$

$$=\sum_{j=1}^{3}\frac{\partial}{\partial x_j}(8\lambda\nu x_i v_{x_j}v_{x_i}) + \frac{\partial}{\partial x_i}(-4\lambda\nu x_i|\bigtriangledown v|^2) + 4\lambda\nu|\bigtriangledown v|^2 - 8\lambda\nu v_{x_i}^2.$$

Hence,

$$8\lambda\nu x_i v_{x_i} \triangle v = 4\lambda\nu|\bigtriangledown v|^2 - 8\lambda\nu v_{x_i}^2 + \sum_{j=1}^{3}\frac{\partial}{\partial x_i}\left(8\lambda\nu x_i v_{x_j}v_{x_i}\right) + \frac{\partial}{\partial x_i}(-4\lambda\nu x_i|\bigtriangledown v|^2).$$

Therefore,

$$2z_1z_2\psi^{\nu+1} = 8\lambda\nu\sum_{i=1}^{3} x_i v_{x_i} \triangle v$$

$$= 4\lambda\nu|\bigtriangledown v|^2 + \sum_{j=1}^{3}\frac{\partial}{\partial x_j}\left(8\lambda\nu v_{x_j}\sum_{i=1}^{3} x_i v_{x_i}\right) + \sum_{i=1}^{3}\frac{\partial}{\partial x_i}(-4\lambda\nu x_i|\bigtriangledown v|^2). \tag{4.4}$$

Second, we estimate $2z_2z_3\psi^{\nu+1}$.

$$2z_2z_3\psi^{\nu+1} = 32\lambda^3\nu^3|x|^2\psi^{-2\nu-2}(1+\gamma)\sum_{i=1}^{3} x_i v_{x_i} v$$

$$=\sum_{i=1}^{3}\frac{\partial}{\partial x_i}\left[16\lambda^3\nu^3\psi^{-2\nu-2}|x|^2(1+\gamma)v^2\right] + 32\lambda^3\nu^3(\nu+1)\psi^{-2\nu-3}|x|^2(1+\gamma)v^2.$$

Therefore, (4.1)–(4.4) and the last inequality imply that

$$(\triangle u)^2\psi^{\nu+1}\alpha^2 \geq 4\lambda\nu|\bigtriangledown v|^2 + 32\lambda^3\nu^3(\nu+1)\psi^{-2\nu-3}|x|^2(1+\gamma)v^2$$

$$+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}\left(8\lambda\nu v_{x_i}\sum_{i=1}^{3}x_i v_{x_i}\right)$$

$$+\sum_{i=1}^{3}\frac{\partial}{\partial x_i}[-4\lambda\nu x_i|\bigtriangledown v|^2 + 16\lambda^3\nu^3\psi^{-2\nu-2}|x|^2(1+\gamma)v^2].$$

Integrating this inequality over Ω and using the fact that $v|_{\partial\Omega}=0$, we obtain, for sufficiently large ν and λ,

$$\int\limits_{\Omega}(\triangle u)^2\psi^{\nu+1}\alpha^2 dx \geq 4\lambda\nu\int\limits_{\Omega}|\bigtriangledown v|^2 dx + 30\lambda^3\nu^3(\nu+1)\int\limits_{\Omega}\psi^{-2\nu-3}|x|^2v^2dx \tag{4.5}$$

$$+\int\limits_{\omega}\left[8\lambda\nu\frac{\partial v}{\partial n}\sum_{i=1}^{3}x_iv_{x_i} - 4\lambda\nu|x|\cos(n,x)|\bigtriangledown v|^2\right]ds.$$

Let $T(x)$ be the tangent plane to ω at the point $x \in \omega$. Introduce the local orthonormal coordinate system with the center at the point $x \in \omega$. The first two coordinate vectors of this system are $s_1(x), s_2(x) \in T(x)$, and the third one is the outward normal vector $n(x)$. Then

$$\sum_{i=1}^{3}x_iv_{x_i}|_\omega = |x|\left[\frac{\partial v}{\partial n}\cos(n,x)+\sum_{i=1}^{3}\frac{\partial v}{\partial s_k}\cos(s_k,x)\right]_\omega,$$

where $\frac{\partial}{\partial s_k}$ is the directional derivative. Because $v|_{\partial\Omega}=0$,

$$\sum_{i=1}^{3}x_iv_{x_i}\;|_\omega = |x|\cos(n,x)\frac{\partial v}{\partial n}\bigg|_\omega. \tag{4.6}$$

Further,

$$|\bigtriangledown v|^2|_\omega = \left[\left(\frac{\partial v}{\partial n}\right)^2 + \sum_{i=1}^{3}\left(\frac{\partial v}{\partial s_k}\right)^2\right]|_\omega = \left(\frac{\partial v}{\partial n}\right)^2|_\omega. \tag{4.7}$$

Hence (4.1), (4.6), and (4.7) imply that

$$\int\limits_{\omega}\left[8\lambda\nu\frac{\partial v}{\partial n}\sum_{i=1}^{3}x_iv_{x_i} - 4\lambda\nu|x|\cos(n,x)|\bigtriangledown v|^2\right]ds$$

$$= 4\lambda\nu\int\limits_{\omega}|x|\cos(n,x)\left(\frac{\partial v}{\partial n}\right)^2ds. \tag{4.8}$$

But since $\cos(n,x)\geq 0$ on $\partial\Omega\backslash\omega_1$ and $\cos(n,x)<0$ on ω_1, (4.8) leads to

$$4\lambda\nu\int\limits_{\omega}|x|\cos(n,x)\left(\frac{\partial v}{\partial n}\right)^2ds \geq 2\lambda\nu\int\limits_{\omega_1}|x|\cos(n,x)\left(\frac{\partial v}{\partial n}\right)^2ds. \tag{4.9}$$

However, since $\frac{\partial v}{\partial n}|_{\omega_1=0}$, the integral in the right-hand side of (4.9) equals zero. Hence, (4.5), (4.8), and (4.9) lead to

$$\int\limits_{\Omega}(\triangle v)^2\psi^{\nu+1}\alpha^2dx \geq C\lambda\nu\int\limits_{\Omega}|\bigtriangledown v|^2\alpha^2dx + C\lambda^3\nu^4\int\limits_{\Omega}v^2\alpha^2dx.$$

Replacing v with u and using the fact that $\nu^4 >> \nu^3$ and $\psi^{\nu+1} < 1$, we obtain

$$\int\limits_{\Omega}(\triangle u)^2\alpha^2dx \geq \int\limits_{\Omega}(\triangle u)^2\psi^{\nu+1}\alpha^2dx$$

$$\geq C\lambda\nu\int\limits_{\Omega}|\bigtriangledown u|^2\alpha^2dx + C\lambda^3\nu^4\int\limits_{\Omega}u^2\alpha^2dx. \tag{4.10}$$

The estimate (4.10), however, does not include the second-order derivatives $u_{x_ix_j}$ in its right-hand side. In order to incorporate them, we refer first to the following classical inequality for the elliptic operators in convex bounded domains Ω (cf. [6]):

$$\int\limits_{\Omega}(\triangle w)^2dx \geq C\int\limits_{\Omega}\sum_{i,j=1}^{3}(w_{x_ix_j})^2dx, \tag{4.11}$$

which is valid for all functions $w \in H^2(\Omega)$ such that $w|_{\omega} = 0$. Further, (4.1) and (4.3) imply that

$$(\triangle u)^2\psi^{\nu+1}\alpha^2 \geq [2z_2(z_1+z_3) + z_1^2 + 2z_1z_3]\psi^{\nu+1}. \tag{4.12}$$

The first bracket term of the right-hand side of (4.12) was already estimated. Below, we will estimate z_1^2 and $2z_1z_3$. To do this, we fix sufficiently large $\nu = \nu_0 = \nu_0(\Omega, \omega_1, x_0)$.

Now we estimate $2z_1z_3\psi^{\nu+1}$. Let $\nu = \nu_0$ and $\lambda \geq \lambda_0$. Similar to the above, we obtain

$$\int\limits_{\Omega}2z_1z_3\psi^{\nu+1}dx \geq -C\lambda^2\int\limits_{\Omega}|\bigtriangledown u|^2\alpha^2dx - C\lambda^3\int\limits_{\Omega}u^2\alpha^2dx. \tag{4.13}$$

Since $\psi \geq \frac{1}{4}$, then $\psi^{\nu+1} = \psi^{\nu_0+1} \geq C$. Hence, (4.11) implies that

$$\int\limits_{\Omega}z_1^2\psi^{\nu+1}dx \geq \left(\frac{1}{4}\right)^{\nu_0+1}\int\limits_{\Omega}(\triangle v)^2dx \geq C\int\limits_{\Omega}\sum_{i,j=1}^{3}(v_{x_ix_j})^2dx.$$

Replacing $v(x)$ in the latter estimate with $u(x)$, we obtain

$$\begin{aligned}\int\limits_{\Omega}z_1^2\psi^{\nu+1}dx \geq C\int\limits_{\Omega}\alpha^2\sum_{i,j=1}^{3}(u_{x_ix_j})^2dx \\ -C\lambda^2\int\limits_{\Omega}|\bigtriangledown u|^2\alpha^2dx - C\lambda^4\int\limits_{\Omega}u^2\alpha^2dx.\end{aligned} \tag{4.14}$$

Further, as we have already proven above, at least

$$\int\limits_{\Omega} 2z_2(z_1+z_3)dx \geq 0.$$

Hence, combining (4.12), (4.13), and (4.14), we obtain

$$\frac{1}{\lambda}\int\limits_{\Omega} (\triangle u)^2\alpha^2 dx \geq \frac{C}{\lambda}\int\limits_{\Omega} \alpha^2 \sum_{i,j=1}^{3} u^2_{x_i x_j} dx$$
$$-C\lambda \int\limits_{\Omega} |\bigtriangledown u|^2\alpha^2 dx - C\lambda^3 \int\limits_{\Omega} u^2\alpha^2 dx.$$

Finally, an obvious combination of the latter estimate with (4.10) leads to

$$\int\limits_{\Omega} (\triangle u)^2\alpha^2 dx \geq \frac{C}{\lambda}\int\limits_{\Omega} \alpha^2 \sum_{i,j=1}^{3} (u_{x_i x_j})^2 dx$$
$$+C\lambda \int\limits_{\Omega} |\bigtriangledown u|^2\alpha^2 dx + C\lambda^3 \int\limits_{\Omega} u^2\alpha^2 dx. \qquad \square$$

5. Discussion. Because of the exponential decay of the Carleman's weight $\alpha(x)$, the major impact in the cost functional J_λ is provided by a small neighborhood $\Omega(x_1) \subset \Omega$ of the point $x_1 \in \omega_1$ closest with x_0. Hence, the data at $\partial\Omega\backslash\omega$ are not really important for J_λ. Therefore, in fact, CWM can work with the backscattering data only. In this case, one needs the Dirichlet data on $\partial\Omega\backslash\omega$ to "bound" the solution. Hence, one does not need to know these data with good precision. By changing x_0, one can cover, by such neighborhoods, a small layer adjacent with ω. Thus, CWM can be considered as a stable layer stripping procedure: $\alpha(x)$ together with the "nonprecise" Dirichlet data at $\partial\Omega\backslash\omega$ provide a stabilization.

Since J_λ is uniformly strictly convex on K_2, then a global convergence on K_2 for a number of standard minimization techniques is guaranteed, which hopefully should lead to an effective numerical implementation(s) of this technique. The crucial point is that in such an implementation, one would not face the problem of local minima.

The major price for these attractive features of CWM consists of predetermination of the number N, which is a regularization parameter. In other words, we work with the Galerkin method without allowing $N \to \infty$. In our opinion, this is acceptable for practical computations. We also note that an increase of N could lead to nonstable algorithms; cf. [1, p. 86] for a similar conclusion. Finally, if one would be able to prove convergence of this method for $N \to \infty$, then one would almost certainly prove a global uniqueness result for this ISP, which is a long-standing problem [4, 8, 9].

REFERENCES

[1] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, New York, 1992.

[2] S. Gutman, M. V. Klibanov, and A. V. Tikhonravov, *Global convexity in a single source 3-D inverse scattering problem*, IMA J. Appl. Math., 55 (1995), pp. 281–302.

[3] L. Hörmander, *Linear Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1966.

[4] V. Isakov, *Inverse Source Problems,* AMS, Providence, RI, 1990.

[5] M. V. Klibanov and O. V. Ioussoupouva, *Uniform strict convexity of a cost functional for 3-D inverse scattering problem*, SIAM J. Math Anal., 26, (1995), pp. 147–179.

[6] O. A. Ladyzenskaya, *Boundary Value Problems of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1985.

[7] M. M. Lavrentiev, V. G. Romanov, and S. P. Shishatskii, *Ill-Posed Problems of Mathematical Physics and Analysis*, AMS, Providence, RI, 1986.

[8] V. G. Romanov, *Inverse Problems of Mathematical Physics*, VNU, Utrecht, The Netherlands, 1987.

[9] A. N. Tikhonov and V. Ya. Arsenin, *Solutions of the Ill-Posed Problems,* Winston-Wiley, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1389–1405, November 1997

007

THE INVERSE CONDUCTIVITY PROBLEM WITH ONE MEASUREMENT: STABILITY AND ESTIMATION OF SIZE*

HYEONBAE KANG[†], JIN KEUN SEO[‡], AND DONGWOO SHEEN[§]

Abstract. We consider the inverse problem to the refraction problem $\mathrm{div}((1+(k-1)\chi_D)\nabla u) = 0$ in Ω and $\frac{\partial u}{\partial \nu} = g$ on $\partial\Omega$. The inverse problem is to determine the size and the location of an unknown object D from the boundary measurement $\Lambda_D(g) = u|_{\partial\Omega}$. The results of this paper are twofold: stability and estimation of size of D. We first obtain upper and lower bounds of the size of D by comparing $\Lambda_D(g)$ with the Dirichlet data corresponding to the harmonic equation with the same Neumann data g. We then obtain logarithmic stability in the case of the disks. In the course of deriving the stability, we are able to compute a positive lower bound (independent of D) of the gradient of the solution u to the refraction problem with the Neumann data g satisfying some mild conditions.

Key words. inverse problem, stability

AMS subject classifications. Primary, 86A20; Secondary, 58G10, 31A25

PII. S0036141096299375

1. Introduction.

Let Ω be a bounded, simply connected domain in $\mathbb{R}^n$ and let D be a subdomain compactly contained in Ω. Let $k \neq 1$ be a positive number and put $\mu = k - 1$. In this paper we consider the inverse problem to the following Neumann problem:

$$P[D,g] \quad \begin{cases} \mathrm{div}((1+\mu\chi_D)\nabla u) = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega, & \displaystyle\int_{\partial\Omega} u = 0, \quad \int_{\partial\Omega} g = 0, \quad g \in L^2(\partial\Omega), \end{cases}$$

where ν is the unit outward normal vector to the boundary $\partial\Omega$. By setting

$$u^e = u|_{\Omega\setminus\overline{D}} \quad \text{and} \quad u^i = u|_D, \tag{1.1}$$

the equation $\mathrm{div}((1+\mu\chi_D)\nabla u) = 0$ may be written as

$$\Delta u^e = 0 \quad \text{in } \Omega \setminus \overline{D}, \tag{1.2}$$

$$\Delta u^i = 0 \quad \text{in } D, \tag{1.3}$$

$$u^e = u^i \quad \text{on } \partial D, \tag{1.4}$$

$$\frac{\partial u^e}{\partial \nu} = k\frac{\partial u^i}{\partial \nu} \quad \text{on } \partial D. \tag{1.5}$$

The inverse problem determines the size and the location of an unknown object D from the boundary measurement $u|_{\partial\Omega} = \Lambda_D(g)$. Practically, it determines D by applying an electric current flux g to $\partial\Omega$ and measures the corresponding electric voltage $\Lambda_D(g)$ on $\partial\Omega$.

*Received by the editors February 26, 1996; accepted for publication (in revised form) August 23, 1996. This research was partially supported by GARC–KOSEF, KOSEF, and BSRI.
http://www.siam.org/journals/sima/28-6/29937.html

[†]Department of Mathematics, Korea University, Seoul 136-701, Korea (kang@semi.korea.ac.kr).
[‡]Department of Mathematics, Yonsei University, Seoul 120-749, Korea (seoj@bubble.yonsei.ac.kr).
[§]Department of Mathematics, Seoul National University, Seoul 151-742, Korea (sheen@math.snu.ac.kr).

Because of its potential of various applications, the inverse conductivity problem has been attracting much attention lately. The uniqueness of the inverse conductivity problem with one (or two) measurement has been studied in many works such as [AIP, BFS, FI, IP, KS, S]. However, there are still plenty of questions to be answered. There are also some works on the stability question [BF, FG, BFI]. The question of the stability estimate is to establish the following kind of inequality:

$$d(D_1, D_2) \leq \psi(\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)}),$$

where $d(D_1, D_2)$ is a distance (e.g., the Hausdorff distance) between D_1 and D_2 and ψ is a function such that $\psi(t) \to 0$ as $t \to 0$. In the paper [BFI], Bellout, Friedman, and Isakov obtained a local stability result with $\psi(t) = Ct$ when $n = 2$ under the assumption that D_1 is a small perturbation of D_2. Without this assumption (i.e., nonlocal), however, no stability estimate for the nonmonotone case has been known.

In this paper, we obtain an estimate of the size of D (in any dimension) without a priori knowledge of the shape of D or and logarithmic stability in the case of two-dimensional disks.

We show that the size of an arbitrary subdomain D of any dimension can be calculated approximately from the boundary measurement (Theorem 3.1). More precisely, we find upper and lower bounds of D in terms of the conductivity k and the quantity $\int_{\partial\Omega}(h - \Lambda_D(g))g d\sigma$, where h is the solution of the Neumann problem $P[\emptyset, g]$, i.e., $\Delta h = 0$ in Ω and $\frac{\partial h}{\partial \nu} = g$ on $\partial\Omega$. In one dimension, where Ω is an interval and D is a subinterval of Ω, one can see via simple computations that

$$\frac{k}{k-1}\int_{\partial\Omega}(h - u) g d\sigma = \|g\|^2_{L^2(\partial\Omega)}|D|.$$

Here, $|D|$ is the Lebesgue size of D. We extend this estimation to higher dimensions in the sense of upper and lower bounds.

We then turn our attention to a theoretical study of the stability in the case of disks of dimension two. In $\mathbb{R}^2$, we suppose that the Neumann data g satisfies the following 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 $\partial\Omega$.)

(N2) $\{P \in \partial\Omega : g(P) \geq 0\}$ and $\{P \in \partial\Omega : g(P) \leq 0\}$ are nonempty connected subsets of $\partial\Omega$.

Let us call these conditions the condition (N). Condition (N1) means that g is rather steep if g is small, while (N2) means that g changes sign only twice on $\partial\Omega$. There are a lot of functions g satisfying the condition (N). For example, when Ω is a disk, $g(P) = \langle \vec{v}, \nu_P \rangle$ with a nonzero constant vector $\vec{v}$ satisfies this condition with $M = |\vec{v}|/2$.

In this paper we find a positive lower bound of $|\nabla u|$ which depends only on the conductivity k and M in (N1) when g satisfies the condition (N) (Theorem 4.1). This is one of the key observations required to derive results from this paper. Based on this observation, we prove that if D_1 and D_2 are disks and $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)}$ is small, then $|D_1 \setminus D_2|$ and $|D_2 \setminus D_1|$ are comparable (see Corollary 4.4). We then prove in Theorem 5.1 the following logarithmic stability for disks:

$$|D_1 \Delta D_2| \leq C \left|\log \|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)}\right|^{-1/\alpha}. \tag{1.6}$$

Here, $|D_1 \Delta D_2|$ is the measure of the symmetric difference of D_1 and D_2 and α is the number determined by the angle between ∂D_1 and ∂D_2 at the intersection points. Let us briefly explain how we obtain (1.6). Let u_j be the solution of $P[D_j, g]$ $(j = 1, 2)$. From the fact that $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)}$ is small, we can show that

$$|u_1 - u_2| \text{ is small in } D_1 \cap D_2, \tag{1.7}$$

$$|\nabla(u_1 - u_2)| \text{ is small outside } D_1 \cup D_2 \tag{1.8}$$

(see Proposition 5.2). We then derive (1.6) using the transmission condition (1.5) on ∂D. Note that in one dimension, where the uniqueness let alone the stability for the inverse conductivity problem does not hold, one can see that (1.7) does not hold while (1.8) is still true.

This paper is organized as follows. In section 2 we review the representation formula for the solution to the refraction problem in [KS]. In section 3 the size estimation of D is derived. In section 4 we prove that $|\nabla u|$ has a positive lower bound if g satisfies the condition (N) and D is a disk. The logarithmic stability of disks is derived in section 5.

The constants C appearing in estimates may vary on each occurrence. However, they are independent of the quantities to be estimated.

2. Representation of the solution to the refraction problem.

In this section we review the representation formula for the refraction problem $P[D, g]$ obtained in [KS]. Let Ω be a simply connected bounded Lipschitz domain in $\mathbb{R}^n$ and let D be a simply connected subdomain with Lipschitz boundary which is compactly contained in Ω. The single layer potential on D is defined by

$$\mathcal{S}_D f(X) = \int_{\partial D} \Gamma(X - Q) f(Q) d\sigma_Q, \qquad X \in \mathbb{R}^n,$$

where $\Gamma(X)$ is the fundamental solution of Δ:

$$\Gamma(X - Q) = \begin{cases} \dfrac{1}{2\pi} \ln|X - Q|, & n = 2, \\ \dfrac{1}{(2-n)\omega_n} |X - Q|^{2-n}, & n \geq 3, \end{cases}$$

and $d\sigma$ is the surface measure. Here ω_n is the area of the unit sphere. Let

$$\mathcal{K}_D^* f(P) = \frac{1}{\omega_n} \int_{\partial D} \frac{\langle \nu_P, P - Q \rangle}{|P - Q|^n} f(Q) d\sigma_Q$$

and $\mathcal{K}_D$ be the dual of $\mathcal{K}_D^*$. The following trace formula is well known (see [F] or [FJR]):

$$\frac{\partial}{\partial \nu} \mathcal{S}_D^{\pm} f(P) := \lim_{t \to 0^+} \langle \nu_P, \nabla \mathcal{S}_D f(P \pm t\nu_P) \rangle = \left(\pm \frac{1}{2} I + \mathcal{K}_D^* \right) f(P) \qquad (P \in \partial D). \tag{2.1}$$

Throughout this paper, $\nabla \mathcal{S}_D^+ f$ and $\nabla \mathcal{S}_D^- f$ denote the restrictions of the gradient of $\mathcal{S}_D f$ to ∂D from the exterior and interior of D, respectively.

Let $L_0^2(\partial\Omega) = \{f \in L^2(\partial\Omega) : \int_{\partial\Omega} f d\sigma = 0\}$. Then the representation formula in [KS] is as follows.

Theorem 2.1 (see [KS]). *If u is a weak solution to the Neumann problem $P[D, g]$, then there are a unique harmonic function $H \in W^{1,2}(\Omega)$ and a density function $\varphi_D \in L_0^2(\partial D)$ such that u can be expressed as*

$$u(X) = H(X) + \mathcal{S}_D \varphi_D(X) \quad \textit{for } X \in \Omega. \tag{2.2}$$

Moreover, if $f = u|_{\partial\Omega}$,

$$H(X) = \mathcal{S}_\Omega \left(-g + \frac{\partial}{\partial\nu} \mathcal{U}_\Omega^+ f|_{\partial\Omega} \right) + c_f \tag{2.3}$$

and

$$\left(\frac{k+1}{2(k-1)} I - \mathcal{K}_D^* \right) \varphi_D = \frac{\partial H}{\partial \nu}|_{\partial D} \qquad \text{on } \partial D, \tag{2.4}$$

where

$$\mathcal{U}_\Omega f(X) = \frac{1}{\omega_n} \int_{\partial\Omega} \frac{\langle Q - X, \nu_Q \rangle}{|X-Q|^n} \left(-\frac{1}{2} I + \mathcal{K}_\Omega \right)^{-1} (f - c_f)(Q) d\sigma_Q, \tag{2.5}$$

$c_f = \int_{\partial\Omega} f \eta_0 d\sigma$, *and* η_0 *is the basis of the null space of* $-\frac{1}{2}I + \mathcal{K}_\Omega^*$ *such that* $\int_{\partial\Omega} \eta_0 d\sigma = 1$.

Note that (2.5) is the solution of the exterior Dirichlet problem $\Delta u = 0$ in $\mathbb{R}^n \setminus \overline{\Omega}$ and $u = f - c_f$ on $\partial\Omega$. Moreover, φ_D satisfies

$$\varphi_D = (k-1) \frac{\partial u^i}{\partial \nu} = \frac{k-1}{k} \frac{\partial u^e}{\partial \nu}. \tag{2.6}$$

When D is a disk, it is known that

$$\mathcal{K}_D^* \varphi(P) = \frac{1}{4\pi r} \int_{\partial D} \varphi(Q) d\sigma_Q \qquad \text{for every } P \in \partial D \tag{2.7}$$

(see [KS]). Therefore, by (2.2), (2.4), and (2.6), we have

$$u = H + \frac{1}{\lambda} \mathcal{S}_D \left(\frac{\partial H}{\partial \nu} \right), \qquad \lambda = \frac{k+1}{2(k-1)}. \tag{2.8}$$

In particular, we have

$$u(X) = \left(1 - \frac{1}{2\lambda} \right) H(X) + \frac{1}{2\lambda} H(X_0), \qquad X \in D, \tag{2.9}$$

where X_0 is the center of D. In fact, by the trace formula (2.1) and (2.7), we have

$$\frac{\partial u}{\partial \nu} = \left(1 - \frac{1}{2\lambda} \right) \frac{\partial H}{\partial \nu}$$

and $u(X_0) = H(X_0)$. Equation (2.9) follows from the uniqueness of the Neumann problem.

3. Size estimations of the unknown object. In this section the domain Ω is contained in $\mathbb{R}^n$ ($n \geq 2$). Let D be a subdomain (not necessarily simply connected) of Ω. Let u_D be the solution of $P[D, g]$, and let h be the solution of $P[\emptyset, g]$, i.e., the harmonic function in Ω with the Neumann boundary data g on $\partial\Omega$ and $\int_{\partial\Omega} h d\sigma = 0$. Put $\mu = k - 1$ as before. For this section we choose the Neumann data g so that the corresponding harmonic function h satisfies

$$\inf_{x \in \Omega} |\nabla h(x)| > 0.$$

There are plenty of Neumann data g which satisfy the above condition: in any dimension, choose, for example, $g(x) = \langle \vec{v}, \nu(x) \rangle$, where $\vec{v}$ is a constant vector. In two dimensions, if g satisfies the condition (N), then the corresponding h satisfies

$$\inf_{x \in \Omega} |\nabla h(x)| > CM,$$

where C is a positive constant depending only on $\partial\Omega$ and the constant M comes from the condition (N1). (See the remark following Theorem 4.1.) The main result of this section is the following theorem.

THEOREM 3.1. *Let u_D and h be as above. Put $C_1 = (\sup_{x \in \Omega} |\nabla h(x)|)^{-2}$ and $C_2 = (\inf_{x \in \Omega} |\nabla h(x)|)^{-2}$. Let*

$$E_D(g) = \left| \int_{\partial\Omega} (h - \Lambda_D(g)) g d\sigma \right|.$$

If $k > 1$, then

$$C_1 \frac{1}{k-1} E_D(g) \le |D| \le C_2 \frac{(\sqrt{k-1}+1)^2}{k-1} E_D(g),$$

and if $0 < k < 1$, then

$$C_1 \frac{(1-\sqrt{1-k})^2}{1-k} E_D(g) \le |D| \le C_2 \frac{1}{1-k} E_D(g).$$

In particular, if $g = \langle \vec{v}, \nu(\cdot) \rangle$, where $\vec{v}$ is an unit constant vector, then $C_1 = C_2 = 1$.

Remark 3.2. Before proving Theorem 3.1, we give an explicit computation of $|D|$ which illustrates the relationship between $|D|$ and $E_D(g)$ when $\Omega = B_1(0)$ and $D = B_d(0)$ in $\mathbb{R}^2$. Here $B_r(x)$ is the disk of radius r centered at x. Let us construct the solution u_D to the refraction problem $P[D, g]$ with $g = \cos\theta$. Using the representation formula (2.8), put

$$u_D = H + \frac{1}{\lambda} \mathcal{S}_D \left(\frac{\partial H}{\partial \nu} \right), \qquad \lambda = \frac{k+1}{2(k-1)},$$

and $H(X) = \alpha x_1 = \alpha r \cos\theta$ (α is a constant to be chosen later). Then, by the trace formula (2.1) and the uniqueness of the Neumann problem, one can see that

$$\frac{1}{\lambda} \mathcal{S}_D \left(\frac{\partial H}{\partial \nu} \right) = \begin{cases} -\dfrac{k-1}{k+1} \alpha r \cos\theta & \text{if } 0 \le r \le d, \\ -\dfrac{k-1}{k+1} \dfrac{\alpha d^2}{r} \cos\theta & \text{if } d < r < 1. \end{cases}$$

Therefore,

$$u_D = \alpha \left(r - \frac{k-1}{k+1} \frac{d^2}{r} \right) \cos\theta \qquad \text{in } \Omega \setminus D.$$

If we choose $\alpha = \frac{k+1}{(k+1)+(k-1)d^2}$, then u_D is the desired solution and

$$\Lambda_D(g) = \frac{(k+1)-(k-1)d^2}{(k+1)+(k-1)d^2} \cos\theta.$$

It is easy to see that $h = r\cos\theta$ is the solution to the harmonic equation with the Neumann data g. Finally, we have

$$\int_{\partial\Omega}(h - \Lambda_D(g))gd\sigma = \frac{2(k-1)}{(k+1)+(k-1)d^2}|D|.$$

LEMMA 3.3. *Let D_j be a domain in Ω and let u_j be the solution of $P[D_j, g]$ $(j = 1, 2)$. Then*

$$\begin{aligned}\int_\Omega (1+\mu\chi_{D_1})|\nabla(u_1-u_2)|^2dx &+ \mu\int_{D_2\setminus D_1}|\nabla u_2|^2dx \\ &= \int_{\partial\Omega}(\Lambda_{D_1}(g) - \Lambda_{D_2}(g))gd\sigma + \mu\int_{D_1\setminus D_2}|\nabla u_2|^2dx.\end{aligned} \tag{3.1}$$

Proof. Since u_1 and u_2 have the same Neumann boundary data g on $\partial\Omega$,

$$\int_\Omega (1+\mu\chi_{D_1})\nabla u_1\nabla\eta dx = \int_\Omega (1+\mu\chi_{D_2})\nabla u_2\nabla\eta dx$$

and hence

$$\int_\Omega (1+\mu\chi_{D_1})\nabla(u_1-u_2)\nabla\eta dx = \mu\int_\Omega (\chi_{D_2}-\chi_{D_1})\nabla u_2\nabla\eta dx \tag{3.2}$$

for every test function $\eta \in W^{1,2}(\Omega)$. Substituting $\eta = u_1$ in (3.2) and integrating by parts, we have

$$\int_{\partial\Omega}(\Lambda_{D_1}(g) - \Lambda_{D_2}(g))gd\sigma = \mu\int_\Omega (\chi_{D_2}-\chi_{D_1})\nabla u_2\nabla u_1 dx, \tag{3.3}$$

while the substitution $\eta = u_1 - u_2$ in (3.2) yields

$$\begin{aligned}&\int_\Omega (1+\mu\chi_{D_1})|\nabla(u_1-u_2)|^2dx \\ &\quad = \mu\int_\Omega (\chi_{D_1}-\chi_{D_2})|\nabla u_2|^2dx + \mu\int_\Omega (\chi_{D_2}-\chi_{D_1})\nabla u_2\nabla u_1 dx.\end{aligned} \tag{3.4}$$

Then Lemma 3.3 follows from (3.3) and (3.4). □

Proof of Theorem 3.1. Note that (3.1) holds even when one of $D_j, j = 1, 2$, is an empty set. Therefore, formula (3.1) with $D_1 = D$ and $D_2 = \emptyset$ leads to

$$\int_\Omega (1+\mu\chi_D)|\nabla(u_D-h)|^2dx - \mu\int_D|\nabla h|^2dx = -\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma, \tag{3.5}$$

and with $D_1 = \emptyset$ and $D_2 = D$ it leads to

$$\int_\Omega |\nabla(u_D-h)|^2dx + \mu\int_D|\nabla u_D|^2dx = \int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma. \tag{3.6}$$

Suppose first that $k > 1$, i.e., $\mu > 0$. Notice that by (3.6),

$$\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma > 0.$$

It then follows from (3.5) that

$$\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma \le \mu\int_D |\nabla h|^2 dx \le \frac{1}{C_1}\mu|D|.$$

On the other hand, by (3.6), we have for any positive ϵ

$$\begin{aligned} |D| &\le C_2\int_D |\nabla h|^2 dx \\ &\le C_2\left[(1+\epsilon)\int_D |\nabla(u_D-h)|^2dx + \left(1+\frac{1}{\epsilon}\right)\int_D |\nabla u_D|^2 dx\right] \\ &\le C_2\alpha(\epsilon;\mu)\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma, \end{aligned}$$

where $\alpha(\epsilon;\mu) = 1+\epsilon+\mu^{-1}(1+\epsilon^{-1})$. $\alpha(\epsilon;\mu)$ has its minimum value when $\epsilon = \frac{1}{\sqrt{\mu}}$, and in this case we have

$$|D| \le C_2\frac{(\sqrt{\mu}+1)^2}{\mu}\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma.$$

We now suppose that $k<1$. Then $-1<\mu<0$. It readily follows from (3.5) that

$$\int_{\partial\Omega}(h-\Lambda_D(g))gd\sigma < 0.$$

We then have from (3.5) that

$$|D| \le C_2\int_D |\nabla h|^2 dx \le C_2\frac{1}{-\mu}\int_{\partial\Omega}(\Lambda_D(g)-h)g\,d\sigma.$$

From (3.6), we see that

$$\int_{\partial\Omega}(\Lambda_D(g)-h)gd\sigma \le -\mu\int_D |\nabla u_D|^2 dx.$$

For every $\epsilon>0$, we obtain from (3.6) that

$$\begin{aligned} \int_D |\nabla u_D|^2 dx &\le (1+\epsilon)\int_D |\nabla(u_D-h)|^2 dx + (1+\epsilon^{-1})\int_D |\nabla h|^2 dx \\ &\le (1+\epsilon)(-\mu)\int_D |\nabla u_D|^2 dx + (1+\epsilon^{-1})\int_D |\nabla h|^2 dx, \end{aligned}$$

which implies that

$$\int_D |\nabla u_D|^2 dx \le \beta(\epsilon;\mu)\int_D |\nabla h|^2 dx,$$

where $\beta(\epsilon;\mu) = \frac{1+\epsilon^{-1}}{1+\mu(1+\epsilon)}$. The minimum of $\beta(\epsilon;\mu)$ with the function value being positive occurs when $\epsilon+1 = 1/\sqrt{-\mu}$, and in this case one has

$$\int_D |\nabla u_D|^2 dx \le \frac{1}{(1-\sqrt{-\mu})^2}\int_D |\nabla h|^2 dx.$$

It then follows that

$$\int_{\partial\Omega}(\Lambda_D(g)-h)gd\sigma \le \frac{-\mu}{(1-\sqrt{-\mu})^2}\int_D |\nabla h|^2 dx \le \frac{1}{C_1}\frac{-\mu}{(1-\sqrt{-\mu})^2}|D|.$$

This completes the proof. □

4. Lower bound of $|\nabla u|$. In this and the next section, D is a *disk* in $\mathbb{R}^2$. In this section, we prove that if u is the solution of $P[D, g]$ where g satisfies the condition (N), then $|\nabla u|$ has a positive lower bound independent of the disk D. We also prove that $|\nabla u|$ has an upper bound independent of D provided that D is at some distance from $\partial\Omega$.

THEOREM 4.1. *Let u be the solution of $P[D, g]$ with g satisfying the condition* (N). *There exists $C_k > 0$ depending only on the conductivity k (particularly independent of D) such that*

$$\inf_{X\in\Omega} |\nabla u(X)| > C_k M.$$

Here M is the number in the condition (N1).

Remark. The above theorem holds even when D is an empty set.

The proof of Theorem 4.1 will be based on the following lemma.

LEMMA 4.2. *Suppose that $D = B_{r_0}(X_0)$. Put $e_1 = (1, 0)$ and $e_2 = (0, 1)$ and let ν be the normal vector field to ∂D. For $j = 1, 2$, let*

$$v_j(X) = \langle e_j, X\rangle + \frac{1}{\lambda}\mathcal{S}_D(\langle e_j, \nu\rangle)(X), \qquad X \in \mathbb{R}^2.$$

Then, v_j $(j = 1, 2)$ satisfies $\operatorname{div}((1 + \mu\chi_D)\nabla v_j) = 0$ *and*

(4.1)
$$\begin{pmatrix}\nabla v_1(X)\\ \nabla v_2(X)\end{pmatrix} = \begin{cases} I - \dfrac{k-1}{k+1} I & \text{if } X \in D,\\ I + \dfrac{k-1}{k+1}\dfrac{r_0^2}{|X - X_0|^2}\begin{pmatrix}\cos 2\theta & \sin 2\theta\\ \sin 2\theta & -\cos 2\theta\end{pmatrix} & \text{if } X \in \mathbb{R}^2 \setminus \overline{D},\end{cases}$$

where I is the identity matrix and $(\sin\theta, \cos\theta) = \frac{X - X_0}{|X-X_0|}$. *In particular,*

$$\left|\begin{matrix}\nabla v_1(X)\\ \nabla v_2(X)\end{matrix}\right| = \begin{cases} \dfrac{4}{(k+1)^2} & \text{if } X \in D,\\ 1 - \left(\dfrac{k-1}{k+1}\dfrac{r_0^2}{|X-X_0|^2}\right)^2 & \text{if } X \in \mathbb{R}^2 \setminus \overline{D},\end{cases}$$

and hence the matrix $(\nabla v_1, \nabla v_2)^t$ is invertible at every point in $\mathbb{R}^2$.

Proof. That v_j satisfies $\operatorname{div}((1 + \mu\chi_D)\nabla v_j) = 0$ is proved in [KS, Lemma 3.3]. Recall that

$$\mathcal{S}_D(\langle e_j, \nu\rangle)(X) = \frac{1}{2\pi}\int_{\partial D} \log|X - Q| \frac{\langle e_j, Q - X_0\rangle}{|Q - X_0|} d\sigma_Q$$

is harmonic in $\mathbb{R}^2 \setminus \partial D$. Put $e(\theta) = X_0 + r_0(\cos\theta, \sin\theta)$. Then, $\nu(e(\theta)) = (\cos\theta, \sin\theta)$ and the tangential vector field $T(e(\theta)) = (-\sin\theta, \cos\theta)$. By the trace formula (2.1) and (2.7), we have

$$\frac{\partial}{\partial\nu}\mathcal{S}_D^{\pm}(\langle e_j, \nu\rangle)(e(\theta)) = \left(\pm\frac{1}{2}I + \mathcal{K}_D^*\right)(\langle e_j, \nu\rangle)(e(\theta)) = \pm\frac{1}{2}\langle e_j, \nu\rangle(e(\theta)). \tag{4.2}$$

It is well known (see [FJR]) that

$$\frac{\partial}{\partial T}\mathcal{S}_D^{+}(\langle e_j, \nu\rangle)(e(\theta)) = \frac{\partial}{\partial T}\mathcal{S}_D^{-}(\langle e_j, \nu\rangle)(e(\theta)). \tag{4.3}$$

It then follows from (4.2) that

$$\nabla \mathcal{S}_D(\langle e_j, \nu\rangle)(X) = -\frac{1}{2} e_j \qquad \text{if } X \in D. \tag{4.4}$$

We also have from (4.3) and (4.4) that

$$\begin{aligned}\nabla \mathcal{S}_D^+(\langle e_1, \nu\rangle)(e(\theta)) &= \langle \nabla \mathcal{S}_D^+(\langle e_1, \nu\rangle), \nu\rangle \nu(e(\theta)) + \langle \nabla \mathcal{S}_D^+(\langle e_1, \nu\rangle), T\rangle T(e(\theta))\\ &= \frac{1}{2}\cos\theta\ \nu(e(\theta)) + \frac{1}{2}\sin\theta\ T(e(\theta))\\ &= \frac{1}{2}(\cos 2\theta, \sin 2\theta).\end{aligned}$$

In the same way, we have

$$\nabla \mathcal{S}_D^+(\langle e_2, \nu\rangle)(e(\theta)) = \frac{1}{2}(\sin 2\theta, -\cos 2\theta).$$

It follows from the uniqueness of the interior and exterior Neumann problem for the harmonic equation that

$$\nabla \mathcal{S}_D(\langle e_1, \nu\rangle)(X) = \begin{cases} -\dfrac{1}{2} e_1 & \text{if } X \in D,\\[2ex] \dfrac{r_0^2}{2|X - X_0|^2}(\cos 2\theta, \sin 2\theta) & \text{if } X \in \mathbb{R}^2 \setminus \overline{D}\end{cases}$$

and

$$\nabla \mathcal{S}_D(\langle e_2, \nu\rangle)(X) = \begin{cases} -\dfrac{1}{2} e_2 & \text{if } X \in D,\\[2ex] \dfrac{r_0^2}{2|X - X_0|^2}(\sin 2\theta, -\cos 2\theta) & \text{if } X \in \mathbb{R}^2 \setminus \overline{D},\end{cases}$$

where $(\cos\theta, \sin\theta) = \frac{X - X_0}{|X - X_0|}$.

Therefore, we have (4.1) and Lemma 4.2 follows. This completes the proof. □

Proof of Theorem 4.1. For a small number $\epsilon > 0$, we can choose $Y_0 \in \Omega \setminus \partial D$ so that

$$|\nabla u(Y_0)| = (1 + \epsilon) \inf_{X \in \Omega \setminus \partial D} |\nabla u(X)|.$$

(Note that by the condition (N2) $\inf_{X \in \Omega\setminus\partial D} |\nabla u(X)| > 0$ (see [AM] or [S]).) By Lemma 4.2, we can choose (a, b) so that

$$\nabla u(Y_0) = a \nabla v_1(Y_0) + b \nabla v_2(Y_0).$$

By (4.1), there is a positive constant C_1 depending only on k such that

$$|(a, b)| \le C_1 |\nabla u(Y_0)|. \tag{4.5}$$

From (4.1) and (4.5), there is a positive constant C_0 depending only on k such that

$$\left| a\frac{\partial v_1}{\partial \nu} + b\frac{\partial v_2}{\partial \nu}\right| \le C_0 |\nabla u(Y_0)|, \tag{4.6}$$

$$\left| \left(a\frac{\partial v_1}{\partial \nu} + b\frac{\partial v_2}{\partial \nu}\right)'\right| \le C_0 |\nabla u(Y_0)| \tag{4.7}$$

for every $X \in \partial\Omega$. Here, $'$ denotes the tangential derivative on $\partial\Omega$. We will show that

$$|\nabla u(Y_0)| > \frac{M}{C_0}.$$

Suppose this is not true. Put

$$\omega(X) = u(X) - av_1(X) - bv_2(X).$$

Then ω satisfies the equation $\operatorname{div}((1+\mu\chi_D)\nabla\omega) = 0$ with the Neumann data

$$\frac{\partial\omega}{\partial\nu} = g - a\frac{\partial v_1}{\partial\nu} - b\frac{\partial v_2}{\partial\nu}.$$

Put

$$I^+ = \{P \in \partial\Omega : g(P) > M\} \quad \text{and} \quad I^- = \{P \in \partial\Omega : g(P) < -M\}.$$

Then, if $P \in \partial\Omega \setminus (I^+ \cup I^-)$, then by the condition $(N1)$, $|g'(P)| > M$ and hence

$$\left|\left(\frac{\partial\omega}{\partial\nu}\right)'(P)\right| \geq |g'(P)| - \left|\left(a\frac{\partial v_1}{\partial\nu} + b\frac{\partial v_2}{\partial\nu}\right)'(P)\right| > M - C_0\frac{M}{C_0} = 0.$$

Therefore, on the set $\partial\Omega \setminus (I^+ \cup I^-)$, $\frac{\partial w}{\partial\nu}$ is strictly monotone. On the other hand, if $P \in I^+$, then

$$\frac{\partial\omega}{\partial\nu}(P) \geq g(P) - \left|\left(a\frac{\partial v_1}{\partial\nu} + b\frac{\partial v_2}{\partial\nu}\right)'(P)\right| > M - C_0\frac{M}{C_0} = 0.$$

Likewise, if $P \in I^-$, then

$$\frac{\partial\omega}{\partial\nu}(P) < 0.$$

Put

$$\partial\Omega \setminus (I^+ \cup I^-) = \cup_{j=1}^m J_j,$$

where J_j are mutually disjoint connected arcs. Since $\frac{\partial\omega}{\partial\nu}$ is monotone on each J_j, if both endpoints of J_j lie in $\overline{I^+}$, then $\frac{\partial\omega}{\partial\nu} > 0$ on J_j. Similarly, if both endpoints of J_j lie in $\overline{I^-}$, then $\frac{\partial\omega}{\partial\nu} < 0$ on J_j. If one endpoint of J_j lies in $\overline{I^+}$ and the other endpoint lies in $\overline{I^-}$, then $\frac{\partial\omega}{\partial\nu}$ changes sign only once on J_j. Since $\{P \in \partial\Omega : g(P) \leq 0\}$ is connected by the condition (N2), we obtain that

$$\left\{P \in \partial\Omega : \frac{\partial\omega}{\partial\nu}(P) \leq 0\right\} \quad \text{is a connected subset of } \partial\Omega.$$

However, since $\nabla\omega(Y_0) = 0$,

$$\left\{P \in \partial\Omega : \frac{\partial\omega}{\partial\nu}(P) \leq 0\right\} \text{ is a disconnected subset of } \partial\Omega.$$

(For this, see [AM].) This gives a contradiction and the proof is complete. □

Although the following lemma is proved in a standard argument, we include a proof for the reader's sake.

LEMMA 4.3. *Let D be a disk in Ω such that $dist(D, \partial\Omega) > \delta$, and let u be the solution to $P[D, g]$. Then there exists $C = C(\delta, k, \Omega, \|g\|_{L^2(\partial\Omega)})$ such that*

$$\sup_{X \in D} |\nabla u(X)| < C.$$

Proof. Let H be the harmonic part of u in the representation (2.2). Then, by (2.9),

$$\nabla u(X) = \left(1 - \frac{1}{2\lambda}\right) \nabla H(X), \qquad X \in D.$$

Since $|\nabla H|$ is subharmonic and $\mathrm{dist}(D, \partial\Omega) > \delta$, we have, for $X \in D$,

$$\begin{aligned} |\nabla H(X)|^2 &\le C_\delta \int_\Omega |\nabla H|^2 dY \\ &= C_\delta \left[- \int_\Omega H \Delta H dY + \int_{\partial\Omega} H \frac{\partial H}{\partial \nu} d\sigma \right] \\ &\le C_\delta \left(\int_{\partial\Omega} \left| \frac{\partial H}{\partial \nu} \right|^2 d\sigma \right)^{1/2} \left(\int_{\partial\Omega} |H|^2 d\sigma \right)^{1/2}. \end{aligned}$$

By (2.1) and (2.3), one sees that

$$\frac{\partial H}{\partial \nu} = \Big(-\frac{1}{2} I + \mathcal{K}_\Omega^* \Big) \Big(-g + \frac{\partial}{\partial \nu} \mathcal{U}_\Omega^+ f|_{\partial\Omega} \Big).$$

Recall that $\mathcal{K}_\Omega^*$ is bounded on $L^2(\partial\Omega)$ and $\mathcal{U}_\Omega$ is bounded on $W^{1,2}(\partial\Omega)$ (see [V]). Moreover, by the Rellich identity,

$$C_1 \left\| \frac{\partial}{\partial T} \mathcal{U}_\Omega f \right\|_{L^2(\partial\Omega)} \le \left\| \frac{\partial}{\partial \nu} \mathcal{U}_\Omega^+ f \right\|_{L^2(\partial\Omega)} \le C_2 \left\| \frac{\partial}{\partial T} \mathcal{U}_\Omega f \right\|_{L^2(\partial\Omega)}$$

for some constant C_1 and C_2 (see [V]). Therefore, we have

$$\left\| \frac{\partial H}{\partial \nu} \right\|_{L^2(\partial\Omega)} \le C \left(\|g\|_{L^2(\partial\Omega)} + \|f\|_{L^2(\partial\Omega)} + \|\frac{\partial f}{\partial T}\|_{L^2(\partial\Omega)} \right).$$

Since $\int_{\partial\Omega} f d\sigma = 0$, it follows from the Poincaré inequality and the Rellich identity again that

$$\|f\|_{L^2(\partial\Omega)} \le C \left\| \frac{\partial f}{\partial T} \right\|_{L^2(\partial\Omega)} \le C \|g\|_{L^2(\partial\Omega)}.$$

It is easy to show that

$$\|H\|_{L^2(\partial\Omega)} \le C \|g\|_{L^2(\partial\Omega)}.$$

In conclusion, we have

$$|\nabla H(X)| \le C(\delta, \Omega, k) \|g\|_{L^2(\partial\Omega)}.$$

This completes the proof. □

As a consequence of Theorem 4.1 and Lemma 4.3, we have the following corollary.

COROLLARY 4.4. *Let D_1 and D_2 be disks in Ω such that dist$(D_j, \partial\Omega) > \delta, j = 1, 2$ for some δ. Then there exist constants C_1 and C_2 such that if $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)} \le \epsilon$, then*

$$C_1(|D_2 \backslash D_1| + \epsilon) \le |D_1 \backslash D_2| + \epsilon \le C_2(|D_2 \backslash D_1| + \epsilon).$$

Proof. Without loss of generality, assume that $k > 1$. By Theorem 4.1, Lemma 4.3, and identity (3.1),

$$|D_2 \setminus D_1| \le C \left(\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)} + |D_1 \setminus D_2| \right).$$

Since (2.1) is symmetric in D_1 and D_2, we simply switch the roles of D_1 and D_2 to obtain the other inequality. This completes the proof. □

5. Stability of disks. In this section we prove logarithmic stability of the disk. Throughout this section let $\Omega_0 = \{X \in \Omega : \text{dist}(X, \partial\Omega) \ge \delta_0\}$ for some fixed $\delta_0 > 0$.

THEOREM 5.1. *Let D_1 and D_2 be disks in Ω_0. Suppose that*

$$|D_1 \cap D_2| \ge \mu_0 \min(|D_1|, |D_2|) \tag{5.1}$$

for some $\mu_0 > 0$. Let g be a Neumann data with the condition (N)*. Then there exists a constant C such that*

$$|D_1 \Delta D_2| \le C \left|\log \|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)}\right|^{-\frac{(3\pi\mu_0)^{1/3}}{\pi}}.$$

Let θ_0 be the angle between ∂D_1 and ∂D_2 at the points of the intersection $\partial D_1 \cap \partial D_2$. Then it is easy to see from elementary geometry of circles that if (5.1) holds, then

$$\theta_0 \ge (3\pi\mu_0)^{1/3}. \tag{5.2}$$

(In fact, assuming that $|D_1| \le |D_2|$, we let $\theta (\le \theta_0)$ be the angle between the tangent line to ∂D_1 at one of the intersection points and the line connecting two intersection points of ∂D_1 and ∂D_2. Then one can see $\theta - \sin\theta \ge \frac{1}{2}\mu_0\pi$.) Therefore, to each point $X \in \Omega \setminus (D_1 \cup D_2)$ there exists a cone $\Gamma(X)$ lying entirely in $\mathbb{R}^2 \setminus (D_1 \cup D_2)$ with the corner at X and the aperture θ_0. Put $\alpha = \pi/\theta_0$.

PROPOSITION 5.2. *Let μ_0 and α be as above. Then there exists a constant C such that if $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)} = \epsilon$, then*

$$\sup_{\Omega_0} |u_1 - u_2| \le C|\log\epsilon|^{-1/\alpha}, \tag{5.3}$$

$$\sup_{\Omega_0 \setminus (D_1 \cup D_2)} |\nabla(u_1^e - u_2^e)| \le C|\log\epsilon|^{-1/\alpha}, \tag{5.4}$$

$$\sup_{D_1 \cap D_2} |\nabla(u_1^i - u_2^i)| \le C\epsilon. \tag{5.5}$$

In order to prove Proposition 5.2, we need the following Lindelöf-type lemma. A similar idea has been used in [A].

Lemma 5.3. *Let z_0 and θ_0 be fixed and let*

$$\Gamma = z_0 + \left\{ re^{i\theta} : 0 < r < 2\rho,\ \theta_0 < \theta < \theta_0 + \frac{\pi}{\alpha} \right\}$$

for some $1 \le \alpha$ and $\rho > 0$. Suppose that ϕ is a holomorphic function in Γ satisfying

$$\begin{cases} \sup\limits_{z \in \Gamma, |z - z_0| \le 2\rho} (|\phi(z)| + |\phi'(z)|) \le L, \\ |\phi(z)| \le \epsilon \quad \text{for all } z \in \Gamma \cap \{\rho \le |z - z_0| \le 2\rho\}. \end{cases}$$

Then there exists C depending only on L such that if $0 \le r < \rho$, then

$$|\phi(z_0 + re^{i\theta})| \le C\epsilon^{(r/\rho)^\alpha \sin \alpha\theta} |\log \epsilon|^{(-1 + (r/\rho)^\alpha)/\alpha}. \tag{5.6}$$

In particular,

$$|\phi(z_0 + re^{i\frac{\pi}{2\alpha}})| \le C\epsilon^{(r/\rho)^\alpha} |\log \epsilon|^{(-1 + (r/\rho)^\alpha)/\alpha}, \tag{5.7}$$

and

$$|\phi(0)| \le C |\log \epsilon|^{-1/\alpha}. \tag{5.8}$$

Proof. Without loss of generality, we may assume that $z_0 = 0$, $\theta_0 = 0$, $\rho = 1$, and $L = 1$. Put $\mathcal{S} = \{re^{i\theta} : 0 < r < 1,\ 0 < \theta < \frac{\pi}{\alpha}\}$. Suppose that w is the harmonic function in Γ with the boundary condition

$$w = 1 \text{ on } \overline{\mathcal{S}} \cap \{|z| = 1\} \quad \text{and} \quad w = 0 \text{ on } \partial\mathcal{S} \setminus \{|z| = 1\}.$$

It follows from the maximum principle that

$$r^\alpha \sin(\alpha\theta) \le w(re^{i\theta}) \le r^\alpha.$$

Note that

$$\log |\phi(z)| \le w(z) \log \epsilon + (1 - w(z)) \log M \quad \text{for all } z \in \partial\mathcal{S},$$

where $M = \sup_{z \in \mathcal{S}} |\phi(z)| \le 1$. By the maximum principle,

$$|\phi(z)| \le \epsilon^{w(z)} M^{1 - w(z)}, \quad z \in \mathcal{S}. \tag{5.9}$$

By the hypothesis, we have

$$|\phi(0)| \le |\phi(re^{i\frac{\pi}{2\alpha}})| + \int_0^r |\phi'(se^{i\frac{\pi}{2\alpha}})| ds \le |\phi(re^{i\frac{\pi}{2\alpha}})| + r,$$

and hence

$$|\phi(0)| \le \epsilon^{r^\alpha} + r \quad \text{for all } 0 < r < 1.$$

By choosing $r^\alpha = \frac{\log |\log \epsilon|}{|\log \epsilon|}$, we get

$$|\phi(0)| \le C |\log \epsilon|^{-1/\alpha}.$$

For any point $z \in \partial\Gamma$ with $|z| \le 1$, we can repeat the above argument with the subsector $\mathcal{S}_z \subset \Gamma$ with the corner at z and congruent to $\mathcal{S}$ to obtain

$$|\phi(z)| \le C|\log \epsilon|^{-1/\alpha}.$$

Therefore,

$$M \le C|\log \epsilon|^{-1/\alpha}.$$

The desired estimates follow from (5.9) and the proof is complete. □

Proof of Proposition 5.2. By the representation formula (2.8), $u_j = H_j + \mathcal{S}_j\varphi_j$ $(j = 1, 2)$, where $\varphi_j = \frac{1}{\lambda}\frac{\partial H_j}{\partial \nu}|_{\partial D_j}$. If we put $f_j = \Lambda_{D_1}(g)$, it follows from (2.3) and a standard argument (see [V]) that

$$\begin{aligned}\|H_1 - H_2\|_{L^2(\partial\Omega)} &= \left\| S_\Omega \left(\frac{\partial}{\partial\nu} U_\Omega(\Lambda_{D_1}(g) - \Lambda_{D_2}(g)) \right) \right\|_{L^2(\partial\Omega)} + C|c_{f_1} - c_{f_2}| \\ &\le C(\|U_\Omega(\Lambda_{D_1}(g) - \Lambda_{D_2}(g))\|_{L^2(\partial\Omega)} + \epsilon) \\ &\le C(\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)} + \epsilon) \\ &\le C\epsilon,\end{aligned}$$

where C depends only on the Lipschitz character of $\partial\Omega$. By standard interior estimates, we have

$$\|H_1 - H_2\|_{L^\infty(\Omega_0)} + \|\nabla(H_1 - H_2)\|_{L^\infty(\Omega_0)} \le C\epsilon. \tag{5.10}$$

So, to prove (5.3), it suffices to show that

$$\|\mathcal{S}_1\varphi_1 - \mathcal{S}_2\varphi_2\|_{L^\infty(\Omega_0)} \le C|\log \epsilon|^{-1/\alpha}. \tag{5.11}$$

Set $W = \mathcal{S}_1\varphi_1 - \mathcal{S}_2\varphi_2$. Then $W \in C(\mathbb{R}^2)$ and it is easy to check that W satisfies the following:

$$\Delta W = 0 \quad \text{in } \mathbb{R}^2 \setminus (\partial D_1 \cup \partial D_2), \tag{5.12}$$

$$\|W\|_{L^2(\partial\Omega)} \le \|H_1 - H_2\|_{L^2(\partial\Omega)} + \|u_1 - u_2\|_{L^2(\partial\Omega)} \le C\epsilon, \tag{5.13}$$

$$|W(X)| + |X||\nabla W(X)| = O(|X|^{-1}) \quad \text{as } |X| \to \infty. \tag{5.14}$$

By the representation formulas (2.8) and (2.9) for the disk, we have

$$|\nabla W(X)| = \frac{1}{2|\lambda|}|\nabla(H_1 - H_2)(X)| \le C\epsilon \quad \text{for every } X \in D_1 \cap D_2. \tag{5.15}$$

Equation (5.5) follows from (5.10) and (5.15). We also have from (2.3) that

$$\sup_{\mathbb{R}^2 \setminus \overline{D_1 \cup D_2}} (|W| + |\nabla W| + |\nabla\nabla W|) \le C(\|H_1\|_{L^2(\Omega)} + \|H_2\|_{L^2(\Omega)}) \le C. \tag{5.16}$$

Since

$$\int_{\mathbb{R}^2 \setminus \Omega} |\nabla W|^2 \le \int_{\partial\Omega} |W||\nabla W| \le C\epsilon$$

by (5.12)–(5.14), we have from the standard interior estimate that

$$\sup_{\{X \in \mathbb{R}^2 : \text{dist}(X,\Omega) \ge 1\}} (|W(X)| + |\nabla W(X)|) \le C\epsilon. \tag{5.17}$$

Therefore, (5.11) follows from the maximum principle, (5.12), (5.15), and the following inequality, still to be proved:

$$\sup_{P\in\partial(D_1\cup D_2)} |W(P)| \le C|\log\epsilon|^{-1/\alpha}. \tag{5.18}$$

To prove (5.18), for each $P \in \partial(D_1 \cup D_2)$, let $\Gamma(P)$ be a cone lying entirely in $\mathbb{R}^2 \setminus D_1 \cup D_2$ with the corner at P and the aperture greater than θ_0 given in (5.2). Let P_1 be a point in $\Gamma(P)$ such that the line joining P and P_1 bisects $\Gamma(P)$ and $\mathrm{dist}(P_1, \Omega) \ge 1$. We claim that

$$|\nabla W(tP + (1-t)P_1)| \le C\epsilon^{t^\alpha}|\log\epsilon|^{(-1+t^\alpha)/\alpha}, \qquad 0 < t < 1. \tag{5.19}$$

In fact, $\phi := W_x - iW_y$ is a holomorphic function in the cone $\Gamma(P)$, $\|\phi\|_{C^1} \le C$ on the bounded subset of $\Gamma(P)$ by (5.16), and, by (5.17),

$$\sup_{X\in\Gamma(P),\mathrm{dist}(X,\Omega)\ge 1} |\phi(X)| \le C\epsilon. \tag{5.20}$$

Thus (5.19) follows from (5.7). By (5.17) and (5.19) we have

$$\begin{aligned}
|W(P)| &\le |W(P_1)| + \int_0^1 \left|\frac{\partial}{\partial t}W(tP + (1-t)P_1)\right| dt \\
&\le C\epsilon + |P - P_1| \int_0^1 |\nabla W(tP + (1-t)P_1)| dt \\
&\le C\epsilon + |\log\epsilon|^{-1/\alpha} \int_0^1 \epsilon^{t^\alpha}|\log\epsilon|^{t^\alpha/\alpha} dt \\
&\le C|\log\epsilon|^{-1/\alpha}.
\end{aligned}$$

This proves (5.18).

Equation (5.8) also says that

$$\sup_{P\in\partial(D_1\cup D_2)} |\phi(P)| \le C|\log\epsilon|^{-1/\alpha}. \tag{5.21}$$

Thus, (5.4) follows from (5.20), (5.21), and the maximum principle. This completes the proof. □

Proof of Theorem 5.1. Put $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^2(\partial\Omega)} = \epsilon$. Then it follows from (5.4) and the transmission condition (1.5) that

$$\begin{aligned}
\frac{\partial}{\partial\nu}(u_1^i - u_2^e) &= \frac{\partial u_1^i}{\partial\nu} - \frac{\partial u_1^e}{\partial\nu} + \frac{\partial}{\partial\nu}(u_1^e - u_2^e) \\
&= (1-k)\frac{\partial u_1^i}{\partial\nu} + O(|\log\epsilon|^{-1/\alpha}) \quad \text{on } \partial D_1 \setminus D_2.
\end{aligned} \tag{5.22}$$

We also have from (5.5) and (1.5) that

$$\begin{aligned}
\frac{\partial}{\partial\nu}(u_1^i - u_2^e) &= \frac{\partial u_1^i}{\partial\nu} - k\frac{\partial u_1^i}{\partial\nu} + k\frac{\partial}{\partial\nu}(u_1^i - u_2^i) \\
&= (1-k)\frac{\partial u_1^i}{\partial\nu} + O(\epsilon) \quad \text{on } \partial D_2 \cap D_1,
\end{aligned} \tag{5.23}$$

where ν is the outward normal vector on $\partial(D_1 \setminus D_2)$. By the Green theorem, we have

$$\int_{D_1\setminus D_2} |\nabla(u_1-u_2)|^2 dx = \int_{\partial(D_1\setminus D_2)} \frac{\partial}{\partial\nu}(u_1^i - u_2^e)(u_1-u_2)d\sigma. \tag{5.24}$$

Thus by (5.3), (5.22), and (5.23),

$$\int_{D_1\setminus D_2} |\nabla(u_1-u_2)|^2 dx = (1-k)\int_{\partial(D_1\setminus D_2)} \frac{\partial u_1^i}{\partial\nu}(u_1-u_2)d\sigma + O(|\log\epsilon|^{-2/\alpha}).$$

Integrating by parts and using (5.22) and (5.23) again, we have

$$\begin{aligned}
&\int_{D_1\setminus D_2} |\nabla(u_1-u_2)|^2 dx \\
&= (1-k)\int_{\partial(D_1\setminus D_2)} u_1\frac{\partial}{\partial\nu}(u_1^i - u_2^e)d\sigma + O(|\log\epsilon|^{-2/\alpha}) \\
&= (1-k)^2\int_{\partial(D_1\setminus D_2)} u_1\frac{\partial u_1^i}{\partial\nu}d\sigma + O(|\log\epsilon|^{-1/\alpha}) \\
&= (1-k)^2\int_{D_1\setminus D_2} |\nabla u_1|^2 dx + O(|\log\epsilon|^{-1/\alpha}).
\end{aligned}$$

By interchanging the role of u_1 and u_2 and adding up, we have

$$\begin{aligned}
&\int_{D_1\setminus D_2} |\nabla u_1|^2 dx + \int_{D_2\setminus D_1} |\nabla u_2|^2 dx \\
&\qquad = \frac{1}{(1-k)^2}\int_{D_1\triangle D_2} |\nabla(u_1-u_2)|^2 dx + O(|\log\epsilon|^{-1/\alpha}).
\end{aligned} \tag{5.25}$$

On the other hand, (5.3) and (5.24) show that

$$\int_{D_1\Delta D_2} |\nabla(u_1-u_2)|^2 dx \le C|\log\epsilon|^{-1/\alpha}.$$

Therefore,

$$\int_{D_1\setminus D_2} |\nabla u_1|^2 dx + \int_{D_2\setminus D_1} |\nabla u_2|^2 dx = O(|\log\epsilon|^{-1/\alpha}).$$

Since $|\nabla u_1|$ and $|\nabla u_2|$ have a lower bound by Theorem 4.1, we have Theorem 5.1. This completes the proof. □

Remark. One of the main ingredients in obtaining the stability result of this paper is the estimate of the lower bound of the solution of $P[D,g]$ (see Theorem 4.1). It would be interesting to see if the same result holds in three dimensions.

REFERENCES

[A] G. Alessandrini, *Stable determination of a crack from boundary measurements*, Proc. Royal Soc. Edinburgh Ser. A, 123 (1993), pp. 497–516.

[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.

[AM] G. Alessandrini and R. Magnanini, *The index of isolated critical points and solutions of elliptic equations in the plane*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 567–589.

[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.

[F] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[FG] A. Friedman and B. Gustafsson, *Identification of conductivity coefficient in an elliptic equation*, SIAM J. Math. Anal., 18 (1987), pp. 777–787.

[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.

[IP] V. Isakov and J. Powell, *On the inverse conductivity problem with one measurement*,Inverse Problems, 6 (1990), pp. 311–318.

[KS] H. Kang and J. K. Seo, *Layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[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. 28, No. 6, pp. 1406–1419, November 1997

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF TWO-DIMENSIONAL PERIODIC SCATTERING PROBLEMS IN ELECTROMAGNETICS[*]

URVE KANGRO[†] AND ROY NICOLAIDES[†]

Abstract. In this paper, two-dimensional electromagnetic scattering problems with a time-periodic incident field are considered. In the case of a perfect conductor scatterer with the presence of an artificial boundary, the existence of a time-periodic solution is proved. For arbitrary initial conditions, asymptotic behavior of solutions is characterized. The asymptotic solution can be represented as the time-periodic solution of the periodically forced scattering problem plus a stationary field. The source of the stationary field is explained, and equations describing it are obtained.

Key words. Maxwell equations, periodic solutions, magnetic offset, electric offset, artificial boundary

AMS subject classifications. 35B40, 35B10, 35Q60

PII. S0036141096298084

1. Introduction.

For solving scattering problems with a time-periodic incident field, the periodic solution is often obtained by introducing an artificial boundary (to limit the domain of computations), choosing arbitrary initial conditions and time-marching Maxwell's curl equations to a periodic state (see [2], [8]). One might expect that as time increases, the solution approaches the solution of the periodically forced scattering problem (the existence of which is proved in section 5). Actually, the difference between these two solutions approaches a steady state, which in general may be nonzero. For the transverse magnetic problem, the electric field is shown to converge to the expected periodic solution, while the magnetic field may have a spurious stationary component; for the transverse electric problem the situation is the opposite. We present here, with complete proofs, the results in case of the transverse magnetic problem, including the equations and the boundary conditions describing the stationary field. The transverse electric problem is briefly discussed in section 10.

Information about the spurious fields is important in practical computations since the computed fields often will be incorrect. To obtain the correct solution of the periodically forced scattering problem, some form of postprocessing is often used. In [6] we discuss some alternatives to postprocessing which enable us to get rid of the spurious stationary fields and at the same time increase the rate of convergence of the numerical methods used. The algorithms in [6] rely on the theorems presented here.

A related problem is considered in [1]. There, authors discuss the decay of solutions for a scatterer with homogeneous boundary conditions and a first-order radiation condition applied at a finite distance from the scatterer. We will solve a similar problem (see sections 6–9), but our method differs somewhat from that of [1]. For instance, we treat Lipschitz boundaries. Also, we do not require some physical compatibility conditions (e.g., the divergence condition) to be satisfied, because they might be difficult to enforce in actual calculations.

[*] Received by the editors January 31, 1996; accepted for publication (in revised form) July 30, 1996. This research was supported by Air Force Office of Scientific Research grant F-49620-94-1-0311.

http://www.siam.org/journals/sima/28-6/29808.html

[†] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (uk22@ andrew.cmu.edu, rn0m@andrew.cmu.edu).

Classical work on the decay of solutions in the acoustic case (in an exterior domain) can be found in [5, Chapter 5].

2. The transverse magnetic problem. Let the scatterer be an infinitely long cylindrical perfect conductor with the axis parallel to the z-axis. We assume that the incoming field has an $\mathbf{E}$-component in the z-direction and an $\mathbf{H}$-component on the xy-plane, both independent of z. Then the scattered field satisfies the same conditions. The corresponding two-dimensional problem is called the transverse magnetic problem.

Let Γ_S be the boundary of the scatterer (on xy-plane), and let Γ be the artificial boundary. We denote the domain between these boundaries by Ω; $\mathbf{n}$ denotes the unit outer normal to Ω. Let $\mathbf{E}_i$ be the incoming electric field. Maxwell's equations for the scattered field are

$$\left\{\begin{array}{lcl} \varepsilon\dfrac{\partial \mathbf{E}}{\partial t} & = & \mathbf{curl}\ \mathbf{H} \\ \mu\dfrac{\partial \mathbf{H}}{\partial t} & = & -\mathbf{curl}\,\mathbf{E} \end{array}\right. \quad \text{in } \Omega, \tag{1}$$

with the boundary conditions

$$\begin{array}{ll} \mathbf{E} = -\mathbf{E}_i & \text{on } \Gamma_S, \\ \mathbf{E} - c\mu\mathbf{H}\times\mathbf{n} = 0 & \text{on } \Gamma, \end{array} \tag{2}$$

and the initial conditions

$$\mathbf{E}(\mathbf{x},0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x},0) = \mathbf{H}_0(\mathbf{x}) \tag{3}$$

(here $\mathbf{x} = (x,y)$). We ignore the divergence conditions, since they are automatically satisfied if the initial conditions are divergence-free. Otherwise, the (nonzero) divergence is preserved in the evolution by equations (1). Since it causes no difficulties in the analysis, we prefer not to make additional assumptions about the initial conditions.

We assume that the incident field can be written in the form $\mathbf{E}_i(\mathbf{x},t) = e^{i\omega t}\widetilde{\mathbf{E}}_i(\mathbf{x})$. The initial conditions for the physical problem typically are not known, and only the periodic solution of (1), (2) of the form $\mathbf{E}(\mathbf{x},t) = e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $\mathbf{H}(\mathbf{x},t) = e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ is of interest. To obtain the periodic solution, arbitrary initial conditions are often used, and the problem is solved on a time interval long enough to make the solution close to a periodic function. We will show below that the limiting solution, while time periodic, is generally not of the form $\mathbf{E}(\mathbf{x},t) = e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $\mathbf{H}(\mathbf{x},t) = e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ but contains an additional static magnetic field, which is caused by "incorrect" (not satisfying certain compatibility conditions) initial conditions.

3. Preliminaries. In this section we will introduce the function spaces and a Green's formula needed later. All functions in the following are complex valued. Define the spaces

$$H^1_{0_S}(\Omega) = \{v \in H^1(\Omega) \mid v|_{\Gamma_S} = 0\}$$

equipped with the norm $|v|_1 = \|\nabla v\|_{L^2}$ (by Poincaré's inequality it is equivalent to the usual H^1-norm) and

$$H_{\mathbf{curl}}(\Omega) = \{\mathbf{u} \in L^2(\Omega)^2 \mid \mathbf{curl}\ \mathbf{u} \in L^2(\Omega)\}$$

with the norm

$$\|\mathbf{u}\|^2_{\mathbf{curl}} = \|\mathbf{u}\|^2_{L^2} + \|\mathbf{curl}\,\mathbf{u}\|^2_{L^2}.$$

We treat the functions with values in $\mathbf{C}^2$ as taking values in $\mathbf{C}^3$ where the third component is zero and the functions with values in $\mathbf{C}$ as having the first two components as zero. If Ω is bounded and the boundary of Ω is Lipschitz continuous, then the unit outer normal $\mathbf{n}$ to its boundary is defined almost everywhere; for $\mathbf{u} \in H_{\mathbf{curl}}(\Omega)$ the trace on the boundary of Ω of $\mathbf{u} \times \mathbf{n}$ is well defined and belongs to $H^{-1/2}(\Gamma \bigcup \Gamma_S)^2$; for $\mathbf{u} \in L^2(\Omega)^2$ with div $\mathbf{u} \in \text{Ł}^2(\Omega)$, the trace of $\mathbf{u} \cdot \mathbf{n}$ is in $H^{-1/2}(\Gamma \bigcup \Gamma_S)$ (see [3] for the trace theorems). The following Green's formula holds:

$$\int_\Omega (\mathbf{u} \cdot \mathbf{curl}\,\mathbf{v} - \mathbf{curl}\,\mathbf{u} \cdot \mathbf{v}) = \langle \mathbf{u} \times \mathbf{n}, \mathbf{v} \rangle_\Gamma \quad \forall \mathbf{u} \in H_{\mathbf{curl}}(\Omega) \text{ and } \forall \mathbf{v} \in H^1_{0_S}(\Omega), \tag{4}$$

where by the dot product on the left-hand side we mean the scalar product in the complex vector space and $\langle \cdot, \cdot \rangle_\Gamma$, defined in $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$, is the extension of the $L^2(\Gamma)$-scalar product in the sense that for $w \in L^2(\Gamma)$ we have

$$\langle w, v \rangle_\Gamma = \int_\Gamma w\overline{v} \quad \forall v \in H^{1/2}(\Gamma).$$

In $H^{1/2}(\Gamma)$ we will use the norm

$$\|\mathbf{f}\|_{H^{1/2}(\Gamma)} = \inf_{\mathbf{u} \in H^1_{0_S}(\Omega),\ \mathbf{u}|_\Gamma = \mathbf{f}} |\mathbf{u}|_1,$$

and in $H^{-1/2}(\Gamma)$ we will use the corresponding dual norm.

4. The main result. In the subsequent sections we will prove the following theorem.

THEOREM 4.1. *Assume that Ω is a bounded multiply-connected domain with a Lipschitz-continuous boundary. Let Γ be the exterior part of the boundary and let Γ_S be the interior boundary (which may consist of several pieces). Assume that the incident field $\mathbf{E}_i$ can be written in the form*

$$\mathbf{E}_i(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{E}}_i(\mathbf{x}),$$

with $\widetilde{\mathbf{E}}_i \in H^2(\Omega)$ and real ω; the initial conditions satisfy $\mathbf{E}_0 \in L^2(\Omega)$ and $\mathbf{H}_0 \in L^2(\Omega)^2$. Then the solutions $\mathbf{E}$, $\mathbf{H}$ of (1)–(3) *satisfy*

$$\left.\begin{array}{l} \|\mathbf{H}(t,\cdot) - e^{i\omega t}\widetilde{\mathbf{H}}(\cdot) - \mathbf{H}^*(\cdot)\|_{L^2} \to 0 \\ \|\mathbf{E}(t,\cdot) - e^{i\omega t}\widetilde{\mathbf{E}}(\cdot)\|_{L^2} \to 0 \end{array}\right. \qquad \textit{as } t \to \infty,$$

where $e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ are unique solutions of this form of (1), (2) *(ignoring the initial conditions) and $\mathbf{H}^*$ is the unique solution of*

$$\begin{cases} \mathbf{curl}\,\mathbf{H}^* = 0, \quad \text{div } \mathbf{H}^* = \text{div } \mathbf{H}_0 & \textit{in } \Omega, \\ \mathbf{H}^* \times \mathbf{n} = 0 & \textit{on } \Gamma, \\ \mathbf{H}^* \cdot \mathbf{n} = \mathbf{H}_0 \cdot \mathbf{n} - \dfrac{1}{i\omega\mu}\mathbf{curl}\,\widetilde{\mathbf{E}}_i \cdot \mathbf{n} & \textit{on } \Gamma_S, \\ \displaystyle\int_{\Sigma_j} \mathbf{H}^* \cdot \mathbf{n} = \int_{\Sigma_j} \left(\mathbf{H}_0 \cdot \mathbf{n} - \frac{1}{i\omega\mu}\mathbf{curl}\,\widetilde{\mathbf{E}}_i \cdot \mathbf{n} \right), & j = 1, \ldots, k-1, \end{cases} \tag{5}$$

where k is the number of connectivity components of Γ_S and Σ_j, $j = 1, \dots, k-1$ are smooth lines in Ω joining the different connectivity components of Γ_S such that $\Gamma_S \cup \left(\cup_{j=1}^{k-1} \Sigma_j\right)$ is connected.

Remark. Under the assumptions of Theorem 4.1, the solution of (1)–(3) satisfies $(\mathbf{E}, \mathbf{H}) \in C([0,\infty); L^2(\Omega) \times L^2(\Omega)^2)$; it solves the integrated (in t) form of the equations (1), (2), (3), i.e., $\int_0^t \mathbf{E} \in H^1(\Omega)$, $\int_0^t \mathbf{H} \in H_{\mathbf{curl}}(\Omega)$, and

$$\varepsilon \mathbf{E} = \varepsilon \mathbf{E}_0 + \mathbf{curl} \int_0^t \mathbf{H} \ , \qquad \mu \mathbf{H} = \mu \mathbf{H}_0 - \mathbf{curl} \int_0^t \mathbf{E} \ .$$

Under the additional assumptions on initial conditions $\mathbf{E}_0 \in H^1(\Omega)$, $\mathbf{H}_0 \in H_{\mathbf{curl}}(\Omega)$ with $\mathbf{E}_0 = -\widetilde{\mathbf{E}}_i$ on Γ_S, and $\mathbf{E}_0 - \mathbf{H}_0 \times \mathbf{n} = 0$ on Γ, we have the classical solution of (1)–(3), $(\mathbf{E}, \mathbf{H}) \in C^1([0,\infty); L^2(\Omega) \times L^2(\Omega)^2) \bigcap C([0,\infty); H^1(\Omega) \times H_{\mathbf{curl}}(\Omega))$.

Idea of proof. First we show the existence of a periodic solution of (1), (2) in the form $e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$. Then consider the equations for the differences

$$\mathbf{E}_d(\mathbf{x}, t) = \mathbf{E}(\mathbf{x}, t) - e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x}), \quad \mathbf{H}_d(\mathbf{x}, t) = \mathbf{H}(\mathbf{x}, t) - e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x}).$$

We can set up the problem in the semigroup framework and, using the fact that the electromagnetic energy is nonincreasing along the solutions, show that the solutions of (1)–(3) with the initial conditions $(\mathbf{E}_0 - \widetilde{\mathbf{E}}, \mathbf{H}_0 - \widetilde{\mathbf{H}})$ (considered as functions of t) stay inside a compact subset of $L^2(\Omega) \times L^2(\Omega)^2$. Then every sequence $(\mathbf{E}_d(t_n), \mathbf{H}_d(t_n))_{n=1}^{\infty}$ with $t_n \to \infty$ as $n \to \infty$ must have at least one accumulation point. We show that the set of all those accumulation points (the ω-limit set) consists of only one point $(0, \mathbf{H}^*)$, where $\mathbf{H}^*$ satisfies (5).

5. Existence of a periodic solution.

In this section we will prove the following result.

THEOREM 5.1. *Assume that Ω and $\mathbf{E}_i$ are as in Theorem 4.1. Then Maxwell's equations (1) together with the boundary conditions (2) have a unique time-periodic solution of the form*

$$\mathbf{E}(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x}), \tag{6}$$

where $\widetilde{\mathbf{E}} \in H^1(\Omega)$ and $\widetilde{\mathbf{H}} \in H_{\mathbf{curl}}(\Omega)$ with div $\widetilde{\mathbf{H}} = 0$.

Proof. Substituting (6) into (1), (2), eliminating $\widetilde{\mathbf{H}}$ from the equations, and dividing everything by $e^{i\omega t}$, we get the following equations for $\widetilde{\mathbf{E}}$ (note that $\mathbf{curl}\,\widetilde{\mathbf{E}} \times \mathbf{n} = \partial\widetilde{\mathbf{E}}/\partial\mathbf{n}$):

$$\begin{cases} \triangle\widetilde{\mathbf{E}} + \dfrac{\omega^2}{c^2}\widetilde{\mathbf{E}} = 0 & \text{in } \Omega, \\ \widetilde{\mathbf{E}} = -\widetilde{\mathbf{E}}_i & \text{on } \Gamma_S, \\ \dfrac{i\omega}{c}\widetilde{\mathbf{E}} + \dfrac{\partial\widetilde{\mathbf{E}}}{\partial\mathbf{n}} = 0 & \text{on } \Gamma, \end{cases} \tag{7}$$

where $c = 1/\sqrt{\varepsilon\mu}$ is the speed of light.

Since the inhomogeneous Dirichlet condition on Γ_S is not easy to handle directly, we transform the problem into a more suitable form. Let ψ be a function in $H^2(\Omega)$ satisfying $\psi = -\widetilde{\mathbf{E}}_i$ on Γ_S and $\psi \equiv 0$ in a neighborhood of Γ. Put $f = -\triangle\psi - (\omega^2/c^2)\psi$.

Then $u = \widetilde{\mathbf{E}} - \psi$ satisfies

$$\begin{cases} \triangle u + \dfrac{\omega^2}{c^2} u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_S, \\ \dfrac{i\omega}{c} u + \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \Gamma. \end{cases} \tag{8}$$

To show the existence of the solution for these equations, we will first use the Lax–Milgram lemma for the equations without the term $(\omega^2/c^2)u$ to define a compact operator B (the inverse of $\triangle$ with the above boundary conditions) and then use the Fredholm alternative to invert $B + \lambda I$. Therefore, let us consider the equations

$$\begin{cases} \triangle u = g & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_S, \\ \dfrac{i\omega}{c} u + \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \Gamma, \end{cases} \tag{9}$$

with $g \in L^2(\Omega)$. The corresponding weak form asks us to find $u \in H^1_{0_S}(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v + \frac{i\omega}{c} \int_\Gamma u \cdot v = -\int_\Omega g \cdot v \quad \forall v \in H^1_{0_S}(\Omega).$$

By the Lax–Milgram lemma (for complex Hilbert spaces) for any $g \in L^2(\Omega)$, there exists a unique weak solution u_g; it satisfies $|u_g|_1 \leq \|g\|_{L^2}$. Define a linear continuous operator $B : L^2(\Omega) \to L^2(\Omega)$ by $Bg = u_g$. Since B maps bounded sets of $L^2(\Omega)$ into bounded sets of $H^1_{0_S}(\Omega)$, B is a compact operator in $L^2(\Omega)$.

Now the equations (8) are equivalent to $u + (\omega^2/c^2)Bu = Bf$. By the Fredholm alternative, this equation is uniquely solvable for any right-hand side if and only if the corresponding homogeneous equation $u + (\omega^2/c^2)Bu = 0$ has only the zero solution. The weak form for the homogeneous equation asks us to find $u \in H^1_{0_S}(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v - \frac{\omega^2}{c^2} \int_\Omega u \cdot v + \frac{i\omega}{c} \int_\Gamma u \cdot v = 0 \quad \forall v \in H^1_{0_S}(\Omega).$$

Substituting $v = u$ in the weak form, we find that $\int_\Gamma |u|^2 = 0$; hence $u = 0$ on Γ (recall that the dot product means the scalar product in a complex vector space). So u is a weak solution of

$$\begin{cases} \triangle u + \dfrac{\omega^2}{c^2} u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_S, \\ u = 0, \quad \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \Gamma. \end{cases} \tag{10}$$

The conditions on Γ imply that the extension of u by zero outside Ω and across Γ is also a solution of $\triangle u + (\omega^2/c^2)u = 0$ in the sense of distributions in a larger domain. Since the eigenfunctions of the Laplacian are (real) analytic (see [4, p. 92]), it follows that u must be identically zero.

From above, it follows that (8) is uniquely solvable for any f. Let u be the solution of (8) with $f = \triangle\psi - (\omega^2/c^2)\psi$. Note that $u \in H^1_{0_S}(\Omega)$. Then $\widetilde{\mathbf{E}} = \psi + u$

is the solution of (7) with $\widetilde{\mathbf{E}} \in H^1(\Omega)$ and $\triangle\widetilde{\mathbf{E}} \in L^2(\Omega)$. Put $\widetilde{\mathbf{H}} = -(1/i\omega\mu)\mathbf{curl}\,\widetilde{\mathbf{E}}$. Then $\mathbf{curl}\,\widetilde{\mathbf{H}} \in L^2(\Omega)$ and $\operatorname{div}\widetilde{\mathbf{H}} = 0$, so $\widetilde{\mathbf{H}} \in H_{\mathbf{curl}}(\Omega)$. It is easy to see that

$$\mathbf{E}(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$$

satisfy Maxwell's equations (1) and the boundary conditions (2). Clearly, the solution is unique. □

6. Semigroup formulation. We will continue the proof of Theorem 4.1. In this section we will prove that the problem for the difference between the solution of (1), (2), (3) and the periodic solution $e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ can be set up in the semigroup framework. This will give us the existence and uniqueness of the solution as well as the smoothness mentioned in the remark after Theorem 4.1. In the subsequent sections we are going to use the standard tools in the semigroup theory for studying the asymptotic behavior of the solutions.

We are going to work in the state space

$$X = \{(\mathbf{E}, \mathbf{H}) \in L^2(\Omega) \times L^2(\Omega)^2\}$$

equipped with the norm

$$\|(\mathbf{E}, \mathbf{H})\|_X^2 = \varepsilon\|\mathbf{E}\|_{L^2}^2 + \mu\|\mathbf{H}\|_{L^2}^2,$$

where the corresponding inner product X is a Hilbert space. Define the operator $A : \mathcal{D}(A) \to X$ by

$$\mathcal{D}(A) = \{(\mathbf{E}, \mathbf{H}) \in H^1_{0_S}(\Omega) \times H_{\mathbf{curl}}(\Omega) \mid \mathbf{E} - c\mu\mathbf{H} \times \mathbf{n}|_\Gamma = 0\},$$

$$A(\mathbf{E}, \mathbf{H}) = \left(\frac{1}{\varepsilon}\mathbf{curl}\,\mathbf{H}, -\frac{1}{\mu}\mathbf{curl}\,\mathbf{E}\right).$$

We can write the problem for the differences

$$\mathbf{E}_d(t) = \mathbf{E}(\cdot, t) - e^{i\omega t}\widetilde{\mathbf{E}}, \quad \mathbf{H}_d(t) = \mathbf{H}(\cdot, t) - e^{i\omega t}\widetilde{\mathbf{H}}$$

(considered as functions of t with values in X) in the form

$$\text{(11)} \qquad \begin{cases} (\dot{\mathbf{E}}_d, \dot{\mathbf{H}}_d) = A(\mathbf{E}_d, \mathbf{H}_d), \\ \mathbf{E}_d(0) = \mathbf{E}_0 - \widetilde{\mathbf{E}}, \quad \mathbf{H}_d(0) = \mathbf{H}_0 - \widetilde{\mathbf{H}}. \end{cases}$$

We want to show that A generates a C_0-contraction semigroup. For this we will use the Lumer–Phillips theorem for Hilbert spaces (see [7]): *Let X be a Hilbert space*; *let $A : \mathcal{D}(A) \to X$ be a linear operator satisfying*

(i) *Re*$\langle Au, u\rangle_X \le 0$ *for all* $u \in \mathcal{D}(A)$,

(ii) $I - A$ *is surjective.*

Then A generates a linear C_0-contraction semigroup.

The first condition (energy estimate) follows from Green's formula (4):

$$\begin{aligned} \text{Re } \langle A(\mathbf{E}, \mathbf{H}), (\mathbf{E}, \mathbf{H})\rangle_X &= \text{Re} \int_\Omega \mathbf{curl}\,\mathbf{H} \cdot \mathbf{E} - \mathbf{curl}\,\mathbf{E} \cdot \mathbf{H} \\ &= -\text{Re } \langle \mathbf{H} \times \mathbf{n}, \mathbf{E}\rangle_\Gamma \\ &= -\frac{1}{c\mu}\int_\Gamma |\mathbf{E}|^2 \quad \forall(\mathbf{E}, \mathbf{H}) \in \mathcal{D}(A). \end{aligned}$$

The second condition needs more work. We have to show that for any $(\mathbf{f},\mathbf{g}) \in X$ we can find $(\mathbf{E},\mathbf{H}) \in \mathcal{D}(A)$ such that

$$\begin{cases} \mathbf{E} - \dfrac{1}{\varepsilon}\mathbf{curl}\,\mathbf{H} = \mathbf{f}, \\ \mathbf{H} + \dfrac{1}{\mu}\mathbf{curl}\,\mathbf{E} = \mathbf{g}. \end{cases}$$

Formally, by eliminating $\mathbf{H}$ we get the following equations for $\mathbf{E}$:

$$\begin{cases} \mathbf{E} + c^2\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} = \mathbf{f} + \dfrac{1}{\varepsilon}\mathbf{curl}\,\mathbf{g} & \text{in } \Omega, \\ \mathbf{E} = 0 & \text{on } \Gamma_S, \\ \mathbf{E} + c\,\mathbf{curl}\,\mathbf{E} \times \mathbf{n} = c\mu\,\mathbf{g} \times \mathbf{n} & \text{on } \Gamma. \end{cases}$$

The corresponding weak form asks us to find $\mathbf{E} \in H^1_{0_S}(\Omega)$ such that

$$(12)\int_\Omega \mathbf{E}\cdot\mathbf{v} + c^2\mathbf{curl}\,\mathbf{E}\cdot\mathbf{curl}\,\mathbf{v} + c\,\langle \mathbf{E},\mathbf{v}\rangle_\Gamma = \int_\Omega \mathbf{f}\cdot\mathbf{v} + \frac{1}{\varepsilon}\mathbf{g}\cdot\mathbf{curl}\,\mathbf{v} \quad \forall \mathbf{v} \in H^1_{0_S}(\Omega).$$

Clearly, the assumptions of the Lax–Milgram lemma are satisfied, so there is a solution of (12), $\mathbf{E} \in H^1_{0_S}(\Omega)$. Put $\mathbf{H} = \mathbf{g} - (1/\mu)\ \mathbf{curl}\,\mathbf{E}$. Then $\mathbf{H} \in L^2(\Omega)^2$ and

$$\mathbf{E} - \frac{1}{\varepsilon}\mathbf{curl}\,\mathbf{H} = \mathbf{E} + c^2\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - \frac{1}{\varepsilon}\mathbf{curl}\,\mathbf{g} = \mathbf{f},$$

hence $\mathbf{H} \in H_{\mathbf{curl}}(\Omega)$. Checking the outer boundary condition, for any $\mathbf{v} \in H^1_{0_S}(\Omega)$ we have

$$\begin{aligned} \langle \mathbf{E} - c\mu\mathbf{H} \times \mathbf{n}, \mathbf{v}\rangle_\Gamma &= \int_\Gamma \mathbf{E}\cdot\mathbf{v}\ + c\mu\int_\Omega \mathbf{curl}\,\mathbf{H}\cdot\mathbf{v} - \mathbf{H}\cdot\mathbf{curl}\,\mathbf{v} \\ &= \int_\Gamma \mathbf{E}\cdot\mathbf{v}\ + c\mu\int_\Omega \varepsilon(\mathbf{E}-\mathbf{f})\cdot\mathbf{v} \\ &\quad -c\mu\int_\Omega \left(\mathbf{g} - \frac{1}{\mu}\mathbf{curl}\,\mathbf{E}\right)\cdot\mathbf{curl}\,\mathbf{v} = 0 \end{aligned}$$

by (12). Consequently, $I - A$ is surjective. From the Lumer–Phillips theorem it follows that A generates a C_0-contraction semigroup $\{T(t) \mid t \geq 0\}$ with $\|T(t)\| \leq 1\ \forall t \geq 0$.

7. Compactness of the orbits. In this section we show that for $(\mathbf{E}^0,\mathbf{H}^0) \in \mathcal{D}(A)$ the solution of (11) stays inside a compact set of X for all $t \geq 0$.

Let $(\mathbf{E}^0,\mathbf{H}^0)$ be given. The orbit through $(\mathbf{E}^0,\mathbf{H}^0)$ is

$$\gamma(\mathbf{E}^0,\mathbf{H}^0) = \{T(t)(\mathbf{E}^0,\mathbf{H}^0) \mid t \geq 0\}.$$

Equip $\mathcal{D}(A)$ with the graph norm

$$\|(\mathbf{E},\mathbf{H})\|_A = \|(\mathbf{E},\mathbf{H})\|_X + \|A(\mathbf{E},\mathbf{H})\|_X.$$

Note that the graph norm is equivalent to the norm in $H^1(\Omega) \times H_{\mathbf{curl}}(\Omega)$. In the following and also in the next section we shall assume that $(\mathbf{E}^0,\mathbf{H}^0) \in \mathcal{D}(A)$. We will remove this assumption in section 9.

Since $\{T(t) \mid t \geq 0\}$ is a contraction semigroup and A commutes with $T(t)$ on $\mathcal{D}(A)$, we have for every $(\mathbf{E}, \mathbf{H}) \in \gamma(\mathbf{E}^0, \mathbf{H}^0)$,

$$\begin{aligned}\|(\mathbf{E}, \mathbf{H})\|_A &= \|T(t)(\mathbf{E}^0, \mathbf{H}^0)\|_X + \|AT(t)(\mathbf{E}^0, \mathbf{H}^0)\|_X \\ &\leq \|(\mathbf{E}^0, \mathbf{H}^0)\|_X + \|A(\mathbf{E}^0, \mathbf{H}^0)\|_X = \|(\mathbf{E}^0, \mathbf{H}^0)\|_A.\end{aligned}$$

Consequently, $\gamma(\mathbf{E}^0, \mathbf{H}^0)$ is bounded in $H^1(\Omega) \times H_{\mathbf{curl}}(\Omega)$.

It is enough to show that the projections of $\gamma(\mathbf{E}^0, \mathbf{H}^0)$ onto $\mathbf{E}$- and $\mathbf{H}$-components have compact closures in $L^2(\Omega)$ and $L^2(\Omega)^2$, respectively. Since the $\mathbf{E}$-component stays bounded in $H^1(\Omega)$ and $H^1(\Omega)$ is compactly embedded into $L^2(\Omega)$, the first compactness is immediate. To show the compactness of $\mathrm{proj}_{\mathbf{H}}\gamma(\mathbf{E}^0, \mathbf{H}^0)$, we note that $\mathbf{E} = 0$ on Γ_S implies that

$$\mathbf{curl}\,\mathbf{E} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_S, \quad \int_{\Sigma_j} \mathbf{curl}\,\mathbf{E} \cdot \mathbf{n} = 0, \quad j = 1, \ldots, k-1,$$

and

$$\frac{d}{dt}\mathrm{div}\ \mathbf{H}(t) = 0 \ \Rightarrow \ \mathrm{div}\ \mathbf{H}(t) = \mathrm{div}\ \mathbf{H}^0.$$

Hence the normal trace of $\mathbf{H} - \mathbf{H}^0$ to Γ_S and to Σ_j makes sense, and

$$(\mathbf{H}(t) - \mathbf{H}^0) \cdot \mathbf{n} = -\frac{1}{\mu}\int_0^t \mathbf{curl}\,\mathbf{E}(\tau) \cdot \mathbf{n}\, d\tau = 0 \quad \text{on } \Gamma_S \tag{13}$$

and

$$\int_{\Sigma_j} (\mathbf{H}(t) - \mathbf{H}^0) \cdot \mathbf{n} = 0, \quad j = 1, \ldots, k-1. \tag{14}$$

Also, on Γ,

$$(\mathbf{H} - \mathbf{H}^0) \times \mathbf{n} = \frac{1}{c\mu}(\mathbf{E} - \mathbf{E}^0) \in H^{1/2}(\Gamma),$$

so the trace of $(\mathbf{H} - \mathbf{H}^0) \times \mathbf{n}$ on Γ is bounded in $H^{1/2}(\Gamma)$. Now let

$$\begin{aligned}S = \{\mathbf{G} \in H_{\mathbf{curl}}(\Omega) | \|\mathbf{G}\|_{\mathbf{curl}} \leq C, \ \|\mathbf{G} \times \mathbf{n}\|_{H^{1/2}(\Gamma)} \leq C, \ \mathrm{div}\ \mathbf{G} = 0 \text{ in } \Omega, \\ \mathbf{G} \cdot \mathbf{n} = 0 \text{ on } \Gamma_S, \ \int_{\Sigma_j} \mathbf{G} \cdot \mathbf{n} = 0, \ j = 1, \ldots, k-1\}.\end{aligned}$$

Note that $\mathbf{H} - \mathbf{H}^0 \in S$. We want to show that S is compact in $L^2(\Omega)^2$.

Suppose $\{\mathbf{G}_k\}_{k=1}^{\infty} \subset S$ is such that $\mathbf{G}_k \rightharpoonup \mathbf{G}$ (weakly) in $H_{\mathbf{curl}}(\Omega)$ as $k \to \infty$. We will show that there is a subsequence $\{\mathbf{G}_{k_l}\}_{l=1}^{\infty}$ such that $\mathbf{G}_{k_l} \to \mathbf{G}$ in $L^2(\Omega)^2$. Indeed, since div $\mathbf{G}_k = 0$ and $\mathbf{G}_k \cdot \mathbf{n} = 0$ on Γ_S, we can choose $\Phi_k \in H^1(\Omega)$ such that $\mathbf{G}_k = \mathbf{curl}\,\Phi_k$. To fix the arbitrary constant in Φ_k, we can require that $\Phi_k = 0$ on Γ_S (since $\mathbf{curl}\,\Phi_k \cdot \mathbf{n} = 0$, Φ_k is constant on each piece of Γ_S; the constants on different pieces of Γ_S must be equal because $\int_{\Sigma_j} \mathbf{curl}\,\Phi \cdot \mathbf{n} = 0$, $j = 1, \ldots, k-1$). The sequence $\{\Phi_k\}_{k=1}^{\infty}$ is bounded in $H^1(\Omega)$ (because $|\Phi_k|_1 = \|\mathbf{G}_k\|_{L^2}$), so there is a subsequence $\{\Phi_{k_l}\}_{l=1}^{\infty}$ converging weakly in $H^1(\Omega)$; the limit Φ of this sequence satisfies $\mathbf{curl}\,\Phi = \mathbf{G}$. Now we have

- $\Phi_{k_l} \to \Phi$ in $L^2(\Omega)$,
- $\Phi_{k_l}|_\Gamma \rightharpoonup \Phi|_\Gamma$ in $H^{1/2}(\Gamma)$,
- $\mathbf{G}_{k_l} \times \mathbf{n} \to \mathbf{G} \times \mathbf{n}$ in $H^{-1/2}(\Gamma)$ (because $H^{1/2}(\Gamma) \hookrightarrow\hookrightarrow H^{-1/2}(\Gamma)$).

Hence by Green's formula (4),

$$\begin{aligned}\int_\Omega \mathbf{G}_{k_l}^2 &= \int_\Omega \mathbf{G}_{k_l} \cdot \mathbf{curl}\ \Phi_{k_l}\\ &= \int_\Omega \mathbf{curl}\,\mathbf{G}_{k_l} \cdot \Phi_{k_l} + \langle \mathbf{G}_{k_l} \times \mathbf{n}, \Phi_{k_l}\rangle_\Gamma + \langle \mathbf{G}_{k_l} \times \mathbf{n}, \Phi_{k_l}\rangle_{\Gamma_S}\\ &\to \int_\Omega \mathbf{curl}\,\mathbf{G} \cdot \Phi + \langle \mathbf{G} \times \mathbf{n}, \Phi\rangle_\Gamma\\ &= \int_\Omega \mathbf{G} \cdot \mathbf{curl}\ \Phi = \int_\Omega \mathbf{G}^2.\end{aligned}$$

Since $\mathbf{G}_{k_l} \rightharpoonup \mathbf{G}$ in $L^2(\Omega)$ and $\|\mathbf{G}_{k_l}\|_{L^2(\Omega)} \to \|\mathbf{G}|_{L^2\Omega)}$ as $l \to \infty$, we have $\mathbf{G}_{k_l} \to \mathbf{G}$ in $L^2(\Omega)$. This implies that the closure of $\mathrm{proj}_\mathbf{H}\gamma(\mathbf{E}^0, \mathbf{H}^0)$ is compact in $L^2(\Omega)^2$. Consequently, the closure of the orbit $\gamma(\mathbf{E}^0, \mathbf{H}^0)$ is compact in X.

8. The $\boldsymbol{\omega}$-limit set. We will study the ω-limit set of $(\mathbf{E}^0, \mathbf{H}^0) \in \mathcal{D}(A)$ and show that it consists of only one point $(0, \mathbf{H}^*)$, where $\mathbf{H}^*$ satisfies (5).

First let us give the definition of the ω-limit set. Let $(\mathbf{E}^0, \mathbf{H}^0) \in X$ be given. The ω-limit set of $(\mathbf{E}^0, \mathbf{H}^0)$, denoted by $\omega(\mathbf{E}^0, \mathbf{H}^0)$, is the set of all $\chi \in X$ such that there exists a sequence of positive times $\{t_n\}_{n=1}^\infty$ with $t_n \to \infty$ as $n \to \infty$ such that $T(t_n)(\mathbf{E}^0, \mathbf{H}^0) \to \chi$ in X.

It is easy to see that if the closure of the orbit through $(\mathbf{E}^0, \mathbf{H}^0)$ is compact, then the ω-limit set $\omega(\mathbf{E}^0, \mathbf{H}^0)$ is nonempty, compact, connected, and invariant (i.e., $T(t)\omega(\mathbf{E}^0, \mathbf{H}^0) \subset \omega(\mathbf{E}^0, \mathbf{H}^0)$ for all $t \geq 0$). Moreover,

$$\mathrm{dist}(T(t)(\mathbf{E}^0, \mathbf{H}^0), \omega(\mathbf{E}^0, \mathbf{H}^0)) \to 0 \text{ as } t \to \infty.$$

Define $V : X \to \mathbb{R}$ by

$$V(\mathbf{E}, \mathbf{H}) = \frac{1}{2}\|(\mathbf{E}, \mathbf{H})\|_X^2.$$

($V(\mathbf{E}, \mathbf{H})$ is the electromagnetic energy.) Since the semigroup is contractive, V is a Liapunov function (nonincreasing along the solutions). Since every element of $\omega(\mathbf{E}^0, \mathbf{H}^0)$ is a limit point of a sequence of the form $T(t_n)(\mathbf{E}^0, \mathbf{H}^0)$, with $t_n \to \infty$ as $n \to \infty$, V must be constant on $\omega(\mathbf{E}^0, \mathbf{H}^0)$. For $(\mathbf{E}^0, \mathbf{H}^0) \in \mathcal{D}(A)$ we can compute

$$\begin{aligned}\frac{d}{dt}V(T(t)(\mathbf{E}^0, \mathbf{H}^0)) &= \mathrm{Re}\ \langle \dot{T}(t)(\mathbf{E}^0, \mathbf{H}^0), T(t)(\mathbf{E}^0, \mathbf{H}^0)\rangle_X\\ &= \mathrm{Re}\ \langle A(\mathbf{E}(t), \mathbf{H}(t)), (\mathbf{E}(t), \mathbf{H}(t))\rangle_X = -\frac{1}{c\mu}\int_\Gamma |\mathbf{E}(t)|^2,\end{aligned}$$

where $(\mathbf{E}(t), \mathbf{H}(t)) = T(t)(\mathbf{E}^0, \mathbf{H}^0)$.

Let $(\mathbf{E}_0^*, \mathbf{H}_0^*) \in \omega(\mathbf{E}^0, \mathbf{H}^0)$ be given. Then $(\mathbf{E}_0^*, \mathbf{H}_0^*) \in \mathcal{D}(A)$ (since $(\mathbf{E}_0^*, \mathbf{H}_0^*)$ is a limit in X of a sequence in $\mathcal{D}(A)$, bounded in the graph norm; there is in $\mathcal{D}(A)$ a weakly convergent subsequence; by the uniqueness of the limit it must converge to $(\mathbf{E}_0^*, \mathbf{H}_0^*)$). Put $(\mathbf{E}^*(t), \mathbf{H}^*(t)) = T(t)(\mathbf{E}_0^*, \mathbf{H}_0^*)$. Since $\omega(\mathbf{E}^0, \mathbf{H}^0)$ is invariant and V is constant on

$\omega(\mathbf{E}^0, \mathbf{H}^0)$, we get that $V(\mathbf{E}^*(t), \mathbf{H}^*(t))$ is constant. Hence $(d/dt)V(\mathbf{E}^*(t), \mathbf{H}^*(t)) = 0$, but this implies that $\mathbf{E}^*(t) = 0$ on Γ for all $t \geq 0$. So $(\mathbf{E}^*(t), \mathbf{H}^*(t))$ are solutions of

$$\tag{15} \begin{cases} \dot{\mathbf{E}}^*(t) = \dfrac{1}{\varepsilon}\mathbf{curl}\,\mathbf{H}^*(t), & \\ \dot{\mathbf{H}}^*(t) = -\dfrac{1}{\mu}\mathbf{curl}\,\mathbf{E}^*(t), & \\ \mathbf{E}^*(t) = 0 & \text{on } \Gamma_S, \\ \mathbf{E}^*(t) = 0, \quad \mathbf{H}^*(t) \times \mathbf{n} = 0 & \text{on } \Gamma, \\ \mathbf{E}^*(0) = \mathbf{E}_0^*, \quad \mathbf{H}^*(0) = \mathbf{H}_0^*. & \end{cases}$$

This system is overdetermined (too many boundary conditions on Γ). We will show that the only $\mathbf{E}^*$ that can satisfy these equations is zero.

Let $\mathbf{u}_k$, $k = 1, 2, \ldots$ be the orthonormal basis of $L^2(\Omega)$ consisting of the eigenfunctions of $-\triangle$ with Dirichlet boundary conditions. Let λ_k, $k = 1, 2, \ldots$ be the corresponding eigenvalues. Expand $\mathbf{E}^*(t)$ in the basis of the eigenfunctions:

$$\mathbf{E}^*(t) = \sum_{k=1}^{\infty} c_k(t)\,\mathbf{u}_k.$$

We will show that $c_k(t) \equiv 0$ for all $k = 1, 2, \ldots$. By taking the inner product of the first equation of (15) with $\mathbf{u}_k$, we get

$$\begin{aligned} \dot{c}_k(t) &= \frac{1}{\varepsilon}\int_\Omega \mathbf{curl}\,\mathbf{H}^*(t) \cdot \mathbf{u}_k = \frac{1}{\varepsilon}\int_\Omega \mathbf{H}^*(t) \cdot \mathbf{curl}\,\mathbf{u}_k \\ &= \frac{1}{\varepsilon}\int_\Omega \mathbf{H}_0^* \cdot \mathbf{curl}\,\mathbf{u}_k - c^2 \int_\Omega \int_0^t \mathbf{curl}\,\mathbf{E}^*(\tau) \cdot \mathbf{curl}\,\mathbf{u}_k \, d\tau \\ &= \frac{1}{\varepsilon}\int_\Omega \mathbf{H}_0^* \cdot \mathbf{curl}\,\mathbf{u}_k - c^2\lambda_k \int_0^t c_k(\tau) d\tau, \end{aligned}$$

so $\ddot{c}_k(t) = -c^2\lambda_k c_k(t)$ and

$$c_k(t) = a_k e^{ic\sqrt{\lambda_k}t} + a_{-k}e^{-ic\sqrt{\lambda_k}t}, \quad k = 1, 2, \ldots.$$

Denote $\gamma_k = c\sqrt{\lambda_k}$, $\gamma_{-k} = -c\sqrt{\lambda_k}$, and $\mathbf{u}_{-k} = \mathbf{u}_k$ for $k = 1, 2, \ldots$. Then we can write

$$\mathbf{E}^*(t) = \sum_{k \neq 0} a_k e^{i\gamma_k t}\mathbf{u}_k.$$

The series is convergent in $H^1(\Omega)$; hence the series $\sum_{k\neq 0} \lambda_k |a_k|^2$ also converges. Now

$$\begin{aligned} \mathbf{H}^*(t) &= \mathbf{H}_0^* - \frac{1}{\mu}\int_0^t \mathbf{curl}\,\mathbf{E}^*(\tau) d\tau \\ &= \mathbf{H}_0^* - \frac{1}{\mu}\int_0^t \sum_{k\neq 0} a_k e^{i\gamma_k \tau}\mathbf{curl}\ \mathbf{u}_k \, d\tau \\ &= \mathbf{H}_0^* - \frac{1}{\mu}\sum_{k\neq 0} \frac{a_k}{i\gamma_k}(e^{i\gamma_k t} - 1)\mathbf{curl}\ \mathbf{u}_k. \end{aligned}$$

Since $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}_k = -\triangle \mathbf{u}_k = \lambda_k \mathbf{u}_k$ and the series $\sum_{k\neq 0} \lambda_k |a_k|^2$ is convergent, the series for $\mathbf{H}^*(t)$ converges in $H_{\mathbf{curl}}(\Omega)$, uniformly in t. So we can calculate the tangential traces on Γ term by term: for any $\mathbf{v} \in H^{1/2}(\Gamma)$ we have

$$0 = \langle \mathbf{H}^*(t) \times \mathbf{n}, \mathbf{v}\rangle_\Gamma = \langle \mathbf{H}_0^* \times \mathbf{n}, \mathbf{v}\rangle_\Gamma - \frac{1}{\mu}\sum_{k\neq 0} \frac{a_k}{i\gamma_k}(e^{i\gamma_k t} - 1)\langle \mathbf{curl}\ \mathbf{u}_k \times \mathbf{n}, \mathbf{v}\rangle_\Gamma .$$

This can be written in the form $\sum_{k\neq 0} b_k e^{i\gamma_k t} =$ const, where we have denoted $b_k = (a_k/\gamma_k)\langle \mathbf{curl}\ \mathbf{u}_k \times \mathbf{n}, \mathbf{v}\rangle_\Gamma$. Note that the convergence of the series is uniform in t. Multiplying the series by $e^{-i\gamma_l t}$ and taking the mean value over $[0,\infty)$, we get (without loss of generality we can assume all γ_k are different)

$$0 = \lim_{T\to\infty} \frac{1}{T}\int_0^T \sum_{k\neq 0} b_k e^{i\gamma_k t} e^{-i\gamma_l t} dt = b_l + \lim_{T\to\infty} \frac{1}{T} \sum_{k\neq 0, k\neq l} \frac{e^{i(\gamma_k - \gamma_l)T} - 1}{i(\gamma_k - \gamma_l)} = b_l .$$

Thus for every k either $a_k = 0$ or $\langle \mathbf{curl}\, \mathbf{u}_k \times \mathbf{n}, \mathbf{v}\rangle_\Gamma = 0$ $\forall \mathbf{v} \in H^{1/2}(\Gamma)$, but the latter implies that $\partial \mathbf{u}_k/\partial \mathbf{n} = 0$ in $H^{-1/2}(\Gamma)$, which contradicts (10) having only the zero solution. Consequently, $a_k = 0$ for all $k \neq 0$; i.e., $\mathbf{E}^*(t) \equiv 0$.

For $\mathbf{H}^*$ we get from (15) first that $\mathbf{H}^*$ is independent of t, and it satisfies

$$\begin{cases} \mathbf{curl}\, \mathbf{H}^* = 0, \\ \mathbf{H}^* \times \mathbf{n} = 0 \quad \text{on } \Gamma. \end{cases} \tag{16}$$

These equations do not determine $\mathbf{H}^*$ uniquely. To get additional equations for $\mathbf{H}^*$, we note that along the orbits

$$\begin{aligned} &\text{div } (\mathbf{H}(t) - \mathbf{H}^0) = 0 && \text{in } \Omega, \\ &(\mathbf{H}(t) - \mathbf{H}^0)\cdot \mathbf{n} = 0 && \text{on } \Gamma_S, \\ &\textstyle\int_{\Sigma_j} (\mathbf{H}(t) - \mathbf{H}^0)\cdot \mathbf{n} = 0, && j = 1,\dots,k-1 \end{aligned} \tag{17}$$

(see (13) and (14)). Since $\mathbf{H}^*$ is the L^2-limit of a sequence in the form $\mathbf{H}(t_n)$ with $t_n \to \infty$ as $n \to \infty$, we get that div $\mathbf{H}^* = 0$. Using the continuity of the trace operator for the normal component from $\{\mathbf{H} \in L^2(\Omega)^2 \,|\, \text{div } \mathbf{H} = 0\}$ to $H^{-1/2}(\Gamma_S)$ and to $H^{-1/2}(\Sigma_j)$, we see that $\mathbf{H}^*$ must also satisfy conditions (17). Now recall that

$$\mathbf{H}^0 = \mathbf{H}_0 - \widetilde{\mathbf{H}} = \mathbf{H}_0 + \frac{1}{i\omega\mu}\mathbf{curl}\, \widetilde{\mathbf{E}}.$$

Since

$$\mathbf{curl}\, \widetilde{\mathbf{E}} \cdot \mathbf{n} = -\mathbf{curl}\, \widetilde{\mathbf{E}}_i \cdot \mathbf{n} \qquad \text{on } \Gamma_S$$

and

$$\int_{\Sigma_j} \mathbf{curl}\, \widetilde{\mathbf{E}} \cdot \mathbf{n} = -\int_{\Sigma_j} \mathbf{curl}\, \widetilde{\mathbf{E}}_i \cdot \mathbf{n}, \qquad j = 1,\dots,k-1,$$

we have

$$(\mathbf{H}^* - \mathbf{H}_0)\cdot \mathbf{n} = -\frac{1}{i\omega\mu}\mathbf{curl}\, \widetilde{\mathbf{E}}_i \cdot \mathbf{n} \qquad \text{on } \Gamma_S$$

and

$$\int_{\Sigma_j} (\mathbf{H}^* - \mathbf{H}_0)\cdot \mathbf{n} = -\frac{1}{i\omega\mu}\int_{\Sigma_j} \mathbf{curl}\, \widetilde{\mathbf{E}}_i \cdot \mathbf{n},$$

which together with (16) gives us (5). Uniqueness of the solution can be shown using the vector potential: the difference between two solutions is $\mathbf{curl}\,\Phi$ for some $\Phi \in H^1_{0_S}(\Omega)$ (note that the conditions on the cuts are needed to find the potential in $H^1_{0_S}(\Omega)$). From the divergence condition we get $\triangle\Phi = 0$ and on the outer boundary Γ we have $\partial\Phi/\partial\mathbf{n} = 0$, hence $\Phi \equiv 0$.

9. General data. In the last two sections we worked under the assumptions that the initial data for the problem for the differences satisfy $(\mathbf{E}^0, \mathbf{H}^0) \in \mathcal{D}(A)$. We want to use the results for problem (11) with $(\mathbf{E}^0, \mathbf{H}^0) = (\mathbf{E}_0 - \widetilde{\mathbf{E}}, \mathbf{H}_0 - \widetilde{\mathbf{H}})$, but even for $(\mathbf{E}_0, \mathbf{H}_0) = (0, 0)$ (which is often the case in practice), we have $(\mathbf{E}^0, \mathbf{H}^0) \notin \mathcal{D}(A)$ since the boundary conditions and the initial conditions are inconsistent. So we must allow $(\mathbf{E}^0, \mathbf{H}^0) \in X$.

Let $(\mathbf{E}^0, \mathbf{H}^0) \in X$ be given. Choose a sequence $(\mathbf{E}^0_k, \mathbf{H}^0_k)^\infty_{k=1}$ in $\mathcal{D}(A)$ such that $(\mathbf{E}^0_k, \mathbf{H}^0_k) \to (\mathbf{E}^0, \mathbf{H}^0)$ in X. Let $\mathbf{H}^*_k$, $k = 1, 2, \ldots$ be the corresponding steady state solutions. Since $T(t)$ is a contraction semigroup, the sequence $(\mathbf{H}^*_k)^\infty_{k=1}$ is a Cauchy sequence in X. Since $\mathbf{curl}\,\mathbf{H}^*_k = 0$, $(\mathbf{H}^*_k)^\infty_{k=1}$ converges in $H_{\mathbf{curl}}(\Omega)$. Let the limit of the sequence be $\mathbf{H}^*$. Then $\mathbf{curl}\,\mathbf{H}^* = 0$, div $\mathbf{H}^* = 0$, and since the corresponding trace operators are continuous in $H_{\mathbf{curl}}(\Omega)$, $\mathbf{H}^*$ also satisfies the boundary conditions in (5). Now for any $k \in \mathbb{N}$ and $t \geq 0$ we have

$$\begin{aligned}\|(\mathbf{E}(t), \mathbf{H}(t) - \mathbf{H}^*)\|_X \leq{}& \|T(t)(\mathbf{E}^0 - \mathbf{E}^0_k, \mathbf{H}^0 - \mathbf{H}^0_k)\|_X \\ &+ \|T(t)(\mathbf{E}^0_k, \mathbf{H}^0_k) - (0, \mathbf{H}^*_k)\|_X + \|\mathbf{H}^*_k - \mathbf{H}^*\|_{L^2(\Omega)}.\end{aligned}$$

By choosing k large we can make the first and the last term arbitrarily small, and for t large the second term is also small. So the results remain true for initial data in X.

This completes the proof of Theorem 4.1.

10. The transverse electric problem. Consider the transverse electric problem; where the scatterer is, as before, an infinitely long cylindrical perfect conductor with the axis parallel to the z-axis but where the incidence field has an $\mathbf{E}$-component in the xy-plane and an $\mathbf{H}$-component on the z-direction. The corresponding scattering problem is

$$\left\{\begin{array}{rcl} \varepsilon\dfrac{\partial\mathbf{E}}{\partial t} &=& \mathbf{curl}\,\mathbf{H} \\ \mu\dfrac{\partial\mathbf{H}}{\partial t} &=& -\mathbf{curl}\ \mathbf{E} \end{array}\right. \quad \text{in } \Omega, \tag{18}$$

with the boundary conditions

$$\begin{array}{ll} \mathbf{E} \times \mathbf{n} = -\mathbf{E}_i \times \mathbf{n} & \text{on } \Gamma_S, \\ \mathbf{E} \times \mathbf{n} + c\mu\mathbf{H} = 0 & \text{on } \Gamma, \end{array} \tag{19}$$

and the initial conditions

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}), \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}_0(\mathbf{x}). \tag{20}$$

In this case one can prove the following theorem.

THEOREM 10.1. *Assume that Ω is a bounded multiply-connected domain with a Lipschitz-continuous boundary. Let Γ be the exterior part of the boundary, and let Γ_S be the interior boundary. Assume that the incident field $\mathbf{E}_i$ can be written in the form*

$$\mathbf{E}_i(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{E}}_i(\mathbf{x}),$$

with $\widetilde{\mathbf{E}}_i \times \mathbf{n} \in H^{-1/2}(\Gamma_S)$ *and real* ω, *and the initial conditions satisfy* $\mathbf{E}_0 \in L^2(\Omega)^2$, $\mathbf{H}_0 \in L^2(\Omega)$. *Then the solutions* $\mathbf{E}$, $\mathbf{H}$ *of* (18)–(20) *satisfy*

$$\left.\begin{array}{l}\|\mathbf{H}(t,\cdot) - e^{i\omega t}\widetilde{\mathbf{H}}(\cdot)\|_{L^2} \to 0 \\ \|\mathbf{E}(t,\cdot) - e^{i\omega t}\widetilde{\mathbf{E}}(\cdot) - \mathbf{E}^*(\cdot)\|_{L^2} \to 0\end{array}\right. \qquad \text{as } t \to \infty,$$

where $e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ *are unique solutions of this form of* (18), (19) *(ignoring the initial conditions), and* $\mathbf{E}^*$ *is the unique solution of*

$$(21) \qquad \begin{cases} \mathbf{curl}\,\mathbf{E}^* = 0, \quad \operatorname{div}\mathbf{E}^* = \operatorname{div}\mathbf{E}_0 & \text{in } \Omega, \\ \mathbf{E}^* \times \mathbf{n} = 0 & \text{on } \Gamma \text{ and on } \Gamma_S, \\ \displaystyle\int_{\Gamma_{S_j}} \mathbf{E}^* \cdot \mathbf{n} = \int_{\Gamma_{S_j}} \mathbf{E}_0 \cdot \mathbf{n}, & j = 1, \dots k, \end{cases}$$

where Γ_{S_j}, $j = 1, \dots k$ *are the connectivity components of* Γ_S.

The idea of proof of the theorem and most of the proof itself are the same (interchanging the roles of $\mathbf{E}$ and $\mathbf{H}$ and using Neumann instead of Dirichlet conditions on Γ_S) as in the case of the transverse magnetic problem, so we will not present it here.

11. Conclusions.

We have shown that the solutions of electromagnetic scattering problems with the incident field of the form $\mathbf{E}_i(\mathbf{x}, t) = e^{i\omega t}\widetilde{\mathbf{E}}_i(\mathbf{x})$ do not always converge to the time-periodic solution of the form $e^{i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$, $e^{i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$. Instead, in case of the transverse magnetic problem there may be a magnetic offset and in case of the transverse electric problem it is possible to have an electric offset. From the equations describing the offset fields ((5) and (21), respectively) we see that if we want the spurious stationary fields to be zero, we have to have additional compatibility conditions on the data. Namely, we have

$$(22) \qquad \operatorname{div}\mathbf{H}_0 = 0 \text{ in } \Omega,$$

$$(23) \qquad \mathbf{H}_0 \cdot \mathbf{n} - \frac{1}{i\omega\mu}\mathbf{curl}\,\widetilde{\mathbf{E}}_i \cdot \mathbf{n} = 0 \text{ on } \Gamma_S,$$

$$(24) \qquad \int_{\Sigma_j} \left(\mathbf{H}_0 \cdot \mathbf{n} - \frac{1}{i\omega\mu}\mathbf{curl}\,\widetilde{\mathbf{E}}_i \cdot \mathbf{n}\right) = 0, \; j = 1, \dots, k-1$$

for the transverse magnetic problem and

$$(25) \qquad \operatorname{div}\mathbf{E}_0 = 0 \text{ in } \Omega,$$

$$(26) \qquad \int_{\Gamma_{S_j}} \mathbf{E}_0 \cdot \mathbf{n} = 0, \; j = 1, \dots k$$

for the transverse electric problem. Condition (23) reflects the requirement that the normal component of the total magnetic field is zero on Γ_S at time $t = 0$ (we can interpret $-\frac{1}{i\omega\mu}e^{i\omega t}\mathbf{curl}\,\widetilde{\mathbf{E}}_i$ as the incident magnetic field at time $t = 0$). If this is not satisfied, then we have a static magnetic field inside the body, which "causes" an additional static magnetic field outside. Condition (24) requires the magnetic flux through surfaces between the scatterers to be zero. Nonzero magnetic flux corresponds to the possibility of currents flowing parallel to the z-axis in the opposite directions in different scatterers (note that the artificial boundary condition implies that the sum of all such currents must be zero in the steady state). Condition (26) means that the total charge on the scatterer(s) must be zero at time $t = 0$. The total charge remains constant in time, and it is this charge which is responsible for the electric offset field.

In practical calculations, it may not be advantageous to impose the compatibility conditions, especially (23), on the initial data. Some alternative methods to eliminate the offset fields are discussed in [6].

For the three-dimensional scattering calculations, similar results are true. Both the electric and magnetic offset may be present, and for simply connected scatterers the equations (5) and (21) describing the offset fields are still true, as well as the compatibility conditions (23) and (26). The main idea of the proof is similar, but one needs additional results about the compactness properties of the spaces related with the operators div and **curl**.

REFERENCES

[1] H. Barucq and B. Hanouzet, *Asymptotic behaviour for Maxwell's system with absorbing boundary conditions*, in Proc. 2nd International Conference on Mathematical and Numerical Aspects of Wave Propagation, R. Kleinmann, T. Angell, D. Colton, F. Santosa, I. Stakgold, eds., Philadelphia, PA, 1993.

[2] C. M. Furse, S. P. Mathur, and O. P. Gandhi, *Improvements to the finite–difference time–domain method for calculating the radar cross section of a perfectly conducting target*, IEEE Trans. Microwave Theory Tech., 38 (1990), p. 919.

[3] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, New York, 1986.

[4] L. Hörmander, *The Analysis of Linear Partial Differential Operators* II, Springer-Verlag, Berlin, New York, 1983.

[5] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

[6] U. Kangro and R.A. Nicolaides, *Spurious fields in time domain computations of scattering problems*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 228–234.

[7] M. Renardy and R.C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1992.

[8] A. Taflove, *Computational Electrodynamics. The Finite Difference Time Domain Method*, Artech House, Boston, MA, 1995.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1420–1445, November 1997

DISCONTINUOUS SOLUTIONS OF A HAMILTON–JACOBI EQUATION WITH INFINITE SPEED OF PROPAGATION[*]

FABIO CAMILLI[†] AND ANTONIO SICONOLFI[†]

Abstract. We consider the Cauchy problem

$$\begin{aligned} u_t - \langle A(x)\nabla u, \nabla u\rangle^{\frac{1}{2}} = 0 \quad & \text{in } \mathbb{R}^N \times (0,+\infty),\\ u = u_0 \quad & \text{in } \mathbb{R}^N \times \{t=0\}, \end{aligned}$$

where $A(x)$ is a positive locally Lipschitz map from $\mathbb{R}^N$ to the space of symmetric matrices. Since no growth condition on A is assumed, pathological phenomena can occur. We study the problem in the framework of discontinuous viscosity solutions and we get some comparison results and a representation formula for the minimal solution of the problem.

Key words. Hamilton–Jacobi equations, discontinuous viscosity solution, Riemannian distance, infinite speed of propagation, canonical solutions

AMS subject classifications. 35F20, 49L25, 35C99

PII. S0036141096298047

1. Introduction.

In this paper we study a Hamilton–Jacobi equation of the type

$$u_t + H(x, \nabla u) = 0 \qquad \text{in } \mathbb{R}^N \times (0,+\infty) \tag{1.1}$$

with the initial condition

$$u = u_0 \qquad \text{in } \mathbb{R}^N \times \{t=0\}, \tag{1.2}$$

where H is a map from $\mathbb{R}^N \times \mathbb{R}^N$ to $\mathbb{R}$ and $u_0 : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is an upper semicontinuous function.

Problem (1.1), (1.2) has been extensively studied in the framework of viscosity solutions theory; see [9], [8], [19]. To get a comparison result for solutions of (1.1), it is often assumed that on H there is a linear growth condition in x. This hypothesis yields a bounded domain of dependence of the solution from the initial datum and therefore the comparison can be carried out into a compact set of $\mathbb{R}^N \times [0,+\infty)$. Roughly speaking, this is like assuming that the first-order PDE (1.1) has the finite propagation speed property.

Here we consider a case where no growth condition on H with respect to x is assumed. More precisely, we take the Hamiltonian as

$$H(x,p) = -\langle A(x)p, p\rangle^{\frac{1}{2}},$$

where A is a locally Lipschitz continuous positive map from $\mathbb{R}^N$ to the space of the symmetric matrices. By means of some examples (see section 8) we show that certain new phenomena can occur in this case. First of all, a solution can be discontinuous

[*] Received by the editors January 29, 1996; accepted for publication (in revised form) August 28, 1996.

http://www.siam.org/journals/sima/28-6/29804.html

[†] Dipartimento di Matematica, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, 00185 Rome, Italy (camilli@axcasp.caspur.it, siconolfi@axcasp.caspur.it).

even starting from a smooth initial datum. To deal with discontinuous viscosity solutions we follow the approach of Barron and Jensen, see [4], [3], exploiting the concavity of H in p, see [17] for an application of this theory to a degenerate problem. Other phenomena are the blow-up of solutions (see Example 8.1) and the lack of uniqueness (see Example 8.2 and Proposition 7.1).

We exhibit a representation formula (see section 5) for a *canonical solution* of problem (1.1), (1.2). To do this we associate to the map A a distance d in $\mathbb{R}^N$ whose properties are described in section 3. The crucial point is that the space $(\mathbb{R}^N, d)$ may not be complete; actually the completeness of this metric is equivalent to the finite propagation speed property for equation (1.1). All the pathologies described above depend on the lack of completeness of $(\mathbb{R}^N, d)$. It is worth noting that the usual condition of linear growth of H in x implies that $(\mathbb{R}^N, d)$ is complete, but the converse is not true as we show in Example 8.3.

One idea essential to the study of the canonical solution is to prove that it is the limit in a suitable weak sense of a sequence of solutions of some approximate problems (see sections 4 and 5).

In section 7 we get some comparison principles which give as a consequence a minimality property of the canonical solution. To prove this we follow the general outline of the analogous results in [4]; we enlarge the set of the test functions as already done in [20] (see Propositions 2.1 and 2.4) and we make use of an explicit locally Lipschitz test function which is constructed in section 6.

Using the comparison principles we are also able to prove that every solution can be represented locally in $\mathbb{R}^N \times (0, +\infty)$ by a formula similar to that of the canonical solution. Furthermore we show that problem (1.1), (1.2) is uniquely solvable once a growth condition at infinity (on the solution) is also prescribed. This is similar to what happens in the parabolic case.

2. Preliminary results.

We now give some notation that we will use in what follows. For any $x = (x^1, \dots, x^N), y = (y^1, \dots, y^N) \in \mathbb{R}^N$, $\mathcal{O} \subset \mathbb{R}^N$ and $r > 0$, we set

$$\langle zx, y\rangle = \sum_{i=1}^{N} x^i y^i, \qquad |x| = \langle x, x\rangle^{\frac{1}{2}},$$

$$B^E(x, r) = \{z \in \mathbb{R}^N : |z - x| < r\}, \qquad d^E(x, \mathcal{O}) = \inf\{|y - x| : y \in \mathcal{O}\}.$$

Let $\mathcal{S}^N$ denote the algebra of the $N \times N$ symmetric matrices endowed with the norm $\|B\| = \max_{|x|=1}\langle Bx, x\rangle$ and the standard order (i.e., for $B, C \in \mathcal{S}^N$, $B \leq C$ if and only if $\langle Bx, x\rangle \leq \langle Cx, x\rangle$ for all $x \in \mathbb{R}^N$). A matrix B is called positive (nonnegative) if $B > 0$ ($B \geq 0$). If $B \geq 0$, then $B^{\frac{1}{2}}$ is the nonnegative matrix verifying $B^{\frac{1}{2}} B^{\frac{1}{2}} = B$. For any $x \in \mathbb{R}^N$, we write $x \otimes x$ for the symmetric matrix with entries $x^i x^j$ for $i, j = 1, \dots, N$. Finally, given $\alpha, \beta \in \mathbb{R}$, $\alpha \wedge \beta$ ($\alpha \vee \beta$) stands for $\min\{\alpha, \beta\}$ ($\max\{\alpha, \beta\}$).

Let us introduce the Cauchy problem we will study:

$$u_t - \langle A(x)\nabla u, \nabla u\rangle^{\frac{1}{2}} = 0 \qquad \text{in } \mathbb{R}^N \times (0, +\infty), \tag{2.1}$$

$$u = u_0 \qquad \text{in } \mathbb{R}^N \times \{t = 0\}, \tag{2.2}$$

where

$$A : \mathbb{R}^N \to \mathcal{S}^N \text{ is locally Lipschitz continuous}, \tag{2.3}$$

$$A(x) > 0 \qquad \text{for all } x \in \mathbb{R}^N, \tag{2.4}$$

and

$$u_0 \in X_0 = \left\{ v : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\} : \ v \text{ is upper semicontinuous} \right\}.$$

From now on we set $H(x,p) = -\langle A(x)p, p\rangle^{\frac{1}{2}}$.

We seek solutions in the functional space

$$X = \left\{ u : \mathbb{R}^N \times [0,+\infty) \to \mathbb{R} \cup \{+\infty\} : \ u \text{ is upper semicontinuous} \right\}.$$

Definition 2.1. *Given an upper semicontinuous (u.s.c.) function $\phi : \mathbb{R}^N \times (0,+\infty) \to \mathbb{R}$ and $(x_0,t_0) \in \mathbb{R}^N \times (0,+\infty)$, we say that a continuous function ψ is supertangent to ϕ at (x_0,t_0) if (x_0,t_0) is a point of local maximum of $\phi - \psi$.*

Definition 2.2. *A function $u \in X$ is said to be a (viscosity) solution of* (2.1) *if for any $(x_0,t_0) \in \mathbb{R}^N \times (0,+\infty)$, $M \in \mathbb{R}$, and for any function $\psi \in C^1(\mathbb{R}^N \times (0,+\infty))$ supertangent to $u \wedge M$ at (x_0,t_0), the following equality holds:*

$$\psi_t(x_0,t_0) + H\left(x_0, \nabla\psi(x_0,t_0)\right) = 0. \tag{2.5}$$

The definition of supersolution (subsolution) is given replacing the equality sign in (2.5) by $\geq$ (resp., $\leq$). As for the initial condition (2.2), it is understood that it has to be taken in the following sense:

$$u_0(x) = \sup\left\{ \limsup_{\varepsilon\to 0} u(x_\varepsilon, t_\varepsilon) : (x_\varepsilon, t_\varepsilon) \ \to (x,0), \ t_\varepsilon > 0 \right\}. \tag{2.6}$$

Exploiting the positive homogeneity of $H(x,p)$ in p, we get the following characterization.

Proposition 2.1.

(i) *A function u is a solution of* (2.1) *if and only if for any (x_0,t_0), $\alpha \leq u(x_0,t_0)$ and for any C^1 function ψ which has a local minimum on the set $\{(x,t) : u(x,t) \geq \alpha\}$ at (x_0,t_0), equality* (2.5) *holds.*

(ii) *A similar equivalence is true for supersolutions and subsolutions.*

Proof. To show that the condition given in the statement implies that u is a solution of (2.1), it is sufficient to note that if $M \in \mathbb{R}$ and ψ is a C^1 supertangent to $u \wedge M$ at (x_0,t_0), then (x_0,t_0) is a point of local minimum of ψ on $\{u \wedge M \geq u(x_0,t_0) \wedge M\} = \{u \geq u(x_0,t_0) \wedge M\}$.

To prove the converse, we use the fact that, due to the positive homogeneity of H in p, (2.1) belongs to the class of geometric equations (see [5], [13], [14]). Then a basic property of such equations permits us to assert that if u is a solution of (2.1), (2.5) holds for any supertangent to $\theta \circ (u \wedge M)$, $M \in \mathbb{R}$ and θ a continuous nondecreasing function.

Now let $\overline{\psi}$ be a C^1 function which attains a local minimum on $\{u \geq \overline{\alpha}\}$ at (x_0,t_0) for a given $\overline{\alpha} \in \mathbb{R}$. Then this point is also a local minimum on $\{u \wedge \overline{M} \geq \overline{\alpha}\}$ for any $\overline{M} \geq \overline{\alpha}$. We refer to Lemma 3.1 of [20] to construct a continuous nondecreasing function $\overline{\theta}$ such that $\overline{\theta} \circ (u \wedge \overline{M})$ permits $\overline{\psi}$ to be a supertangent at (x_0,t_0). This ends the proof of part (i). Minor modifications of the above argument also give the assertion (ii). □

The next step is to enlarge the set of test functions for equation (2.1) in such a way as to include the locally Lipschitz ones. We use an approximation argument by means of *inf convolution.*

Let us recall some properties of the (Clarke) generalized gradient that we will use later (see [6] for a general treatment).

Proposition 2.2. *Let $\phi : \mathbb{R}^N \times [0,+\infty) \to \mathbb{R}$ be a locally Lipschitz continuous function and denote by $\partial\phi(z,t)$ its generalized gradient. Then we have the following:*

(i) *The multifunction $(z,t) \to \partial\phi(z,t)$ is u.s.c.; i.e., for any (z_0,t_0), $(z_n,t_n) \to (z_0,t_0)$, $(q_n,s_n) \in \partial\phi(z_n,t_n)$ with (q_n,s_n) converging to (q,s), one has $(q,s) \in \partial\phi(z_0,t_0)$.*

(ii) *If ϕ attains a local extremum at (z_0,t_0), then $(0,0) \in \partial\phi(z_0,t_0)$.*

(iii) *If ϕ is semiconcave, then*

$$\partial\phi(z,t) = \big\{(q,s) : \textit{there exists } \overline{\phi}\ C^1\textit{-supertangent} \\ \textit{to } \phi \textit{ at } (z,t) \textit{ with } (\nabla\overline{\phi}(z,t), \overline{\phi}_t(z,t)) = (q,s)\big\}.$$

(iv)

$$\partial\phi(z,t) = \mathrm{co}\Big\{ \lim_{n\to\infty} (\nabla\phi(z_n,t_n), \phi_t(z_n,t_n)) : \\ \phi \textit{ is differentiable at } (z_n,t_n) \textit{ and } (z_n,t_n) \to (z,t)\Big\},$$

where co *means convex hull.*

The following proposition summarizes some properties of the inf convolution. We refer to [12], [18] for definitions and basic properties.

Proposition 2.3. *Let $\phi : \mathbb{R}^N \times [0,+\infty) \to \mathbb{R}$ be continuous. For $\lambda > 0$ define*

$$\phi_\lambda(z,t) = \inf_{y,s} \left\{ \phi(y,s) + \frac{\lambda}{2}(|y-z|^2 + |t-s|^2) \right\}. \tag{2.7}$$

Then

(i) *ϕ_λ is semiconcave.*

(ii) *if $(\overline{y},\overline{s})$ realizes the infimum in* (2.7), *then $\lambda(\overline{y}-z, \overline{s}-t) \in \partial\phi_\lambda(z,t)$.*

Proposition 2.4. *Let $(x_0,t_0) \in \mathbb{R}^N \times (0,+\infty)$, $\alpha \in \mathbb{R}$, and ψ be a locally Lipschitz function which attains a local minimum on $\{u \geq \alpha\}$ at (x_0,t_0). Then, if u is a solution of* (2.1), *there exists $(p,s) \in \partial\psi(x_0,t_0)$ such that*

$$s + H(x_0,p) = 0. \tag{2.8}$$

If instead u is a supersolution (subsolution), then

$$s + H(x_0,p) \geq 0 \quad (\leq 0). \tag{2.9}$$

Proof. Suppose first that u is a solution. Regularize ψ using inf convolution by defining the sequence

$$\psi_n(x,t) = \inf_{y,s} \Big\{ \psi(y,s) + \frac{n}{2}(|x-y|^2 + |t-s|^2) \Big\} \tag{2.10}$$

which converges to ψ uniformly on compact subsets of $\mathbb{R}^N \times (0,+\infty)$. Note that the addition to ψ of $-\beta(|x-x_0|^2+|t-t_0|^2)$, $\beta > 0$, does not affect the generalized gradient at (x_0,t_0). Therefore we can suppose without loss of generality that the minimum (x_0,t_0) is strict. Hence straightforward arguments provide a sequence (x_n,t_n) of local minima of ψ_n on the closed set $\{u \geq \alpha\}$ converging, up to a subsequence, to (x_0,t_0).

Now denote by (y_n,r_n) a point satisfying

$$\psi_n(x_n,t_n) = \psi(y_n,r_n) + \frac{n}{2}(|x_n-y_n|^2 + |t_n-r_n|^2)$$

and set

$$p_n = n(y_n - x_n), \qquad s_n = n(r_n - t_n). \tag{2.11}$$

Observe that

$$(p_n, s_n) \in \partial\psi_n(x_n, t_n) \cap \partial\psi(y_n, r_n) \tag{2.12}$$

according to Proposition 2.2(ii) and Proposition 2.3(ii). Exploiting Proposition 2.3(i), Proposition 2.2(iii), and Proposition 2.1, from (2.12) we get

$$s_n + H(x_n, p_n) = 0. \tag{2.13}$$

Moreover, (y_n, r_n) is convergent to (x_0, t_0) and the sequence (p_n, s_n) is bounded by (2.12) and so is convergent up to a subsequence. Denoting the limit by (p, s), we obtain, by Proposition 2.2(i),

$$(p, s) \in \partial\psi(x_0, t_0)$$

and so (2.8) follows letting $n \to +\infty$ in (2.13). Suitable modifications of the above argument also give (2.9). □

We conclude this section with two remarks.

Remark 2.1. Note that since H is nonpositive and because of (2.6), any solution u is nondecreasing for $t \in [0, +\infty)$.

Remark 2.2. If u is a subsolution of (2.1) and ψ is a test function as in Proposition 2.4 (which attains a local minimum at (x_0, t_0) on the intersection $\{u \geq \alpha\} \cap (\mathbb{R}^N \times (0, t_0])$, $\alpha \in \mathbb{R}$), then (2.9) still holds; see [5]. In this case even if u is a solution, one cannot expect equality (2.8) but only (2.9).

3. Properties of the distance function.

Here, starting from any given continuous map from $\mathbb{R}^N$ to $\mathcal{S}^N$ with nonnegative values, we define a distance on $\mathbb{R}^N$. This construction will allow us to give a representation of a *canonical* solution of (2.1), (2.2) as well as of some approximated problems we will introduce in the next section.

Definition 3.1. *Given a continuous map $C : \mathbb{R}^N \to \mathcal{S}^N$ with $C(x) \geq 0$, a continuous piecewise C^1 curve $\xi : [0, T] \to \mathbb{R}^N$ is said to be C-admissible if*

$$\dot\xi(t) \otimes \dot\xi(t) \leq C(\xi(t)) \qquad \textit{a.e. in } [0, T]. \tag{3.1}$$

We set for $x, y \in \mathbb{R}^N$,

$$\begin{aligned} d_C(x, y) = \inf\{T : &\text{ there exists a } \xi\ C\text{-admissible} \\ &\text{with } \xi(0) = x,\ \xi(T) = y\} \end{aligned} \tag{3.2}$$

with the convention that $d_C(x, y) = +\infty$ if there are no C-admissible paths joining x to y. It is an exercise to verify that d_C is a distance on $\mathbb{R}^N$, possibly infinite. Moreover, d_C turns $\mathbb{R}^N$ into a length space in the sense of [15].

We consider the distance d_A associated to the map $A(x)$ which appears in (2.1). Thanks to (2.4), d_Δ is a finite distance. We will hereafter write d instead of d_A and set $B(x, r) = \{y \in \mathbb{R}^N : d(x, y) < r\}$ for any $r > 0$, $x \in \mathbb{R}^N$.

Proposition 3.1.

(i) *d is topologically equivalent to the Euclidean distance.*

(ii) *For any $x \in \mathbb{R}^N$, the set $\{r > 0 : B(x, r)$ is relatively compact$\}$ is a nonempty, open interval.*

(iii) *The metric space* $(\mathbb{R}^N, d)$ *is complete if and only if for every* $x \in \mathbb{R}^N, r > 0$, *the ball* $B(x,r)$ *is relatively compact.*

Note that, because of (i), the expression relatively compact in (ii) and (iii) is not ambiguous.

Proof. We first show that for every x_0, r, the Euclidean ball $B^E(x_0,r)$ is open in the topology induced by d. More precisely we set for every $x \in B^E(x_0,r)$, $T_x = d^E(x, \partial B^E(x_0,r))$, where ∂ denotes the boundary in the Euclidean topology and we claim that

$$B\left(x, \frac{T_x}{M}\right) \subset B^E(x_0,r) \tag{3.3}$$

for any M satisfying

$$A(x) \leq M^2 I, \qquad x \in B^E(x_0,r). \tag{3.4}$$

If (3.3) is not true, we can find $x \in B^E(x_0,r)$, $T > 0$, and an A-admissible path ξ joining $x = \xi(0)$ to a point $\xi(T) \in \partial B^E(x_0,r)$ such that

$$T < \frac{T_x}{M} \tag{3.5}$$

and

$$\xi([0,T)) \subset B^E(x_0,r). \tag{3.6}$$

The last formula, together with (3.4), gives

$$|\dot{\xi}(t)| \leq M \qquad \text{a.e. in [0,T]}.$$

Then we get

$$T_x \leq |x - \xi(T)| \leq \int_0^T |\dot{\xi}(t)| dt \leq MT,$$

which contradicts (3.5).

We conclude the proof of (i) by proving that, for any x_0, r, $B(x_0,r)$ is open in the Euclidean topology. According to the positivity of A (see (2.4)) for any $x \in B(x_0,r)$ and $\delta > 0$, we can choose a positive number $m_\delta = m_\delta(x)$ such that

$$A(y) \geq m_\delta^2 I, \qquad y \in B^E(x,\delta). \tag{3.7}$$

If $y \in B^E(x,\delta)$, then the path ξ defined via

$$\xi(t) = x + m_\delta \frac{y-x}{|y-x|} t, \qquad t \in [0, |y-x|/m_\delta],$$

is A-admissible thanks to the relation

$$\dot{\xi}(t) \otimes \dot{\xi}(t) = m_\delta^2 \left(\frac{y-x}{|y-x|} \otimes \frac{y-x}{|y-x|}\right) \leq m_\delta^2 I \leq A(\xi(t))$$

which holds by (3.7). Hence

$$B^E(x,\delta) \subset B\left(x, \frac{\delta}{m_\delta}\right). \tag{3.8}$$

Since m_δ is nondecreasing with respect to δ, we can select $\overline{\delta}$ so that

$$\frac{\overline{\delta}}{m_{\overline{\delta}}} < r - d(x, x_0). \tag{3.9}$$

Finally we use (3.8), (3.9) to find

$$B^E(x, \overline{\delta}) \subset B(x, r - d(x, x_0)) \subset B(x_0, r),$$

which ends the proof of (i).

Now fix $x \in \mathbb{R}^N$ and denote by I_x the set that appears in (ii). Exploiting (i) and the local compactness of the Euclidean topology, we see that I_x is nonempty. Moreover I_x is obviously an interval. Hence we will prove (ii), once we show that for any given $\delta \in I_x$, there exists $\delta' > \delta$ belonging to I_x. To do this, choose $r > 0$ such that

$$B(x, \delta) \subset B^E\left(x, \frac{r}{2}\right). \tag{3.10}$$

Set $\mathcal{K} = \partial B^E(x, r)$ and define

$$f(y) = d(y, \mathcal{K}) = \inf_{\mathcal{K}} d(y, z)$$

for any $y \in \mathbb{R}^N$. The function f is continuous by (i), so taking into account that $\delta \in I_x$ and (3.10), we have that

$$\sigma = \inf_{B(x,\delta)} \{f\}$$

is positive. Therefore $B(x, \delta + \sigma) \subset B^E(x, r)$, so $\delta + \sigma \in I_x$, which proves (ii).

It is easy to see that the relative compactness of the balls implies the completeness of d. The converse follows from a general theorem for locally compact length space (see [15, Chapter 1]). □

For any $x \in \mathbb{R}^N$, let us denote by $d(x, \infty)$ the supremum of the set $\{r > 0 : B(x, r)$ is relatively compact$\}$. This notation is justified by the fact that whenever $d(x, \infty)$ is finite, it is the smallest radius for which the ball (for the distance d) centered in x becomes unbounded with respect to the Euclidean distance. The completeness of $(\mathbb{R}^N, d)$ is clearly equivalent to $d(x, \infty)$ being infinite for every x. On the other hand, $(\mathbb{R}^N, d)$ is not complete if and only if $d(x, \infty) < +\infty$ for every x.

LEMMA 3.1. *Assume that $(\mathbb{R}^N, d)$ is not complete and let x_0 be any point of $\mathbb{R}^N$. Every sequence contained in $B(x_0, d(x_0, \infty))$ has a subsequence which is fundamental with respect to d.*

Proof. The argument of Theorem 1.10 in [15] easily can be adapted to verify our Lemma 3.1. □

LEMMA 3.2. *Let $x_0, y_0 \in \mathbb{R}^N$ and $r > 0$ such that $\overline{B(x_0, r)}$ is compact and $y_0 \notin \overline{B(x_0, r)}$. Then there exist $z_0, z \in \partial B(x_0, r)$ such that*

$$d(y_0, z_0) = d(x_0, y_0) - r, \tag{3.11}$$

$$d(z, \infty) = d(x, \infty) - r. \tag{3.12}$$

Proof. To show (3.11) we consider, for every $n \in \mathbb{N}$, an A-admissible curve ξ_n verifying $\xi_n(0) = x_0$, $\xi_n(d(x_0, y_0) + \frac{1}{n}) = y_0$ and denote by z_n a point belonging to the intersection of this curve with $\partial B(x_0, r)$. We have

$$d(x_0, y_0) - r \leq d(z_n, y_0) \leq d(x_0, y_0) + \frac{1}{n} - r. \tag{3.13}$$

From (3.13) we see that every z_0, limit point of z_n, satisfies (3.11).

We turn to the proof of (3.12). If $(\mathbb{R}^N, d)$ is complete, then (3.12) is obvious for any z. If $(\mathbb{R}^N, d)$ is not complete, consider a sequence y_n contained in $B(x_0, d(x_0, \infty))$ such that $|y_n| \to +\infty$. By Lemma 3.1, y_n is fundamental for d up to a subsequence. We denote by z_n, for any n, an element of $\partial B(x_0, r)$ satisfying

$$d(z_n, y_n) = d(x_0, y_n) - r < d(x_0, \infty) - r. \tag{3.14}$$

Such a point exists by virtue of (3.11). We can assume without loss of generality that z_n converges to a limit z. Taking into account the properties of y_n and (3.14), we have

$$\begin{aligned} d(z_n, y_m) &\leq d(z_n, y_n) + d(y_n, y_m) \\ &< d(x_0, \infty) - r + \varepsilon \end{aligned} \tag{3.15}$$

for any given $\varepsilon > 0$ and n, m sufficiently large. Letting $n \to +\infty$ in (3.15), we find that the inequality

$$d(z, y_m) < d(x_0, \infty) - r + \varepsilon \tag{3.16}$$

holds for m sufficiently large. Since $|y_m| \to +\infty$, (3.16) implies that $B(z, d(x_0, \infty) - r + \varepsilon)$ is noncompact, so $d(z, \infty) \leq d(x, \infty) - r + \varepsilon$, which gives the assertion since ε is arbitrary. □

Let us stress that our assumptions on A do not exclude the possibility that $(\mathbb{R}^N, d)$ is not complete. We shall prove later that this fact is the cause of the pathologies in the analysis of problem (2.1). To emphasize this, we now show that as a consequence of Proposition 3.1 the growth conditions in x usually assumed on the Hamiltonian (see, for example, [2]) lead to the completeness of $(\mathbb{R}^N, d)$.

Corollary 3.1. *Assume that there exist positive constants α, β such that*

$$|H(x, p)| \leq (\alpha|x| + \beta)|p| \qquad \textit{for every } x, p. \tag{3.17}$$

Then $(\mathbb{R}^N, d)$ is complete.

Proof. Condition (3.17) implies that

$$\|A^{\frac{1}{2}}(x)\| \leq \alpha|x| + \beta, \qquad x \in \mathbb{R}^N.$$

So, if ξ is an A-admissible path, we find

$$|\dot{\xi}(t)| \leq \alpha|\xi(t)| + \beta \qquad \text{for a.e. } t. \tag{3.18}$$

Given $x_0 \in \mathbb{R}^N$, $r > 0$, by Gronwall's inequality, (3.17), and (3.2), we obtain

$$|x| \leq (|x_0| + \beta r)e^{\alpha r}$$

for every $x \in B(x_0, r)$. This shows that $B(x_0, r)$ is relatively compact, so the assertion follows according to Proposition 3.1(iii). □

The converse of the previous corollary is not true as we will show in Example 8.3.

4. Approximation by complete metrics. In this section we approximate A by a sequence of maps $A_n : \mathbb{R}^N \to \mathcal{S}^N$ with compact support. Moreover, we study the distances d_n associated to A_n and define functions w_n which we will prove in the next section to be convergent in a suitable sense to a solution of (2.1), (2.2).

To simplify notation, from now on $\mathcal{O}_n$ will stand for $B^E(0, 2n)$ for every $n \in \mathbb{N}$.

Proposition 4.1. *There exists a sequence of maps A_n from $\mathbb{R}^N$ to $\mathcal{S}^N$ verifying*

$$A_n(x) > 0 \quad \textit{for } x \in \mathcal{O}_n, \qquad A_n = 0 \quad \textit{for } x \in \mathbb{R}^N \setminus \mathcal{O}_n, \tag{4.1}$$

$$A_n \leq A, \tag{4.2}$$

$$A_n,\ A_n^{\frac{1}{2}} \in C^\infty(\mathbb{R}^N, \mathcal{S}^N), \tag{4.3}$$

$$A_n \textit{ converges to } A \textit{ uniformly on compact subsets of } \mathbb{R}^N. \tag{4.4}$$

Proof. Regularize the elements of $A(x)$ via mollification to obtain a sequence $C_n(x)$, with $C_n : \mathbb{R}^N \to \mathcal{S}^N$ smooth, which converges to A uniformly on compact subsets of $\mathbb{R}^N$. Set

$$M_n = \max_{\overline{\mathcal{O}}_n} \|A(x)\|,$$
$$m_n = \min_{\overline{\mathcal{O}}_n} \|A(x)\|,$$

and denote by ε_n a positive sequence verifying

$$\varepsilon_n < \frac{1}{n}\left(\frac{m_n}{M_n}\right) \wedge m_n, \tag{4.5}$$

where M_n/m_n is well defined because m_n is strictly positive by (2.4). From the convergence of C_n, it follows that we can find a subsequence, still denoted by C_n, which verifies

$$A(x) - \varepsilon_n I \leq C_n(x) \leq A(x) + \varepsilon_n I, \qquad x \in \overline{\mathcal{O}}_n. \tag{4.6}$$

Consequently we have

$$C_n(x) > 0, \qquad x \in \overline{\mathcal{O}}_n, \tag{4.7}$$

$$\|C_n(x)\| \leq 2M_n, \qquad x \in \overline{\mathcal{O}}_n. \tag{4.8}$$

Setting $\gamma_n = m_n/(m_n + \varepsilon_n)$ and using (4.5), (4.6), and (4.8), we get for any $x \in \overline{\mathcal{O}}_n$ the following inequalities:

$$\gamma_n C_n(x) \leq \gamma_n A(x) + \varepsilon_n \gamma_n I \leq \frac{m_n}{m_n + \varepsilon_n} A(x) + \frac{\varepsilon_n}{m_n + \varepsilon_n} A(x) = A(x) \tag{4.9}$$

and

$$\begin{aligned} \|\gamma_n C_n(x) - A(x)\| &\leq |\gamma_n - 1| \|C_n(x)\| + \|A(x) - C_n(x)\| \\ &\leq 2\frac{\varepsilon_n}{m_n + \varepsilon_n} M_n + \varepsilon_n \leq \frac{1}{n}\frac{m_n}{M_n}\frac{1}{m_n} 2M_n + \varepsilon_n \leq \frac{3}{n}. \end{aligned} \tag{4.10}$$

Now we consider a C^∞ cut-off function λ defined in $[0, +\infty)$ with the following properties:

$$\lambda(t) = 1, \quad t \in [0, 1]; \qquad 0 < \lambda(t) < 1, \quad t \in [1, 4]; \qquad \lambda(t) = 0, \quad t \in [4, \infty).$$

Further, we set, for any $t \in \mathbb{R}^+$, $x \in \mathbb{R}^N$, $\mu_n(t) = \lambda^2(\frac{t}{n^2})$, and $A_n(x) = \mu_n(|x|^2)\gamma_n C_n(x)$. By (4.7) and (4.9), A_n satisfies (4.1) and (4.2).

To obtain (4.3), note that since C_n is strictly positive, $C_n^{\frac{1}{2}}$ is smooth and the same is true for $A_n^{\frac{1}{2}}(x) = \lambda(\frac{|x|^2}{n^2})\sqrt{\gamma_n} C_n^{\frac{1}{2}}(x)$.

To prove (4.4), observe that, since any compact subset $\mathcal{K} \subset \mathbb{R}^N$ is contained in $B^E(0, n)$ for n sufficiently large, from (4.10) it follows for such n that

$$\max_{\mathcal{K}} \|A_n(x) - A(x)\| \leq \max_{\overline{\mathcal{O}}_n} \|\gamma_n C_n(x) - A(x)\| \leq \frac{3}{n}$$

which ends the proof. □

In connection with the maps A_n, we can define distances d_n as in the previous section, namely, via (3.1) and (3.2) with C replaced by A_n. We set for any $x \in \mathbb{R}^N$ and $r > 0$, $B_n(x, r) = \{y : d_n(x, y) < r\}$.

PROPOSITION 4.2.

(i) *For any* $x_0 \in \mathbb{R}^N \setminus \mathcal{O}_n$, $r > 0$,

$$B_n(x_0, r) = \{x_0\} \tag{4.11}$$

or, equivalently, $d_n(x_0, y) = +\infty$ *if* $y \neq x_0$.

(ii) d_n *is the distance induced on* $\mathcal{O}_n$ *by the Riemannian metric* g_n *defined via*

$$g_n(x) = A_n^{-1}(x), \qquad x \in \mathcal{O}_n. \tag{4.12}$$

Proof. If $x_0 \in \mathbb{R}^N \setminus \overline{\mathcal{O}}_n$, assertion (i) is clear. Consider $x_0 \in \partial\mathcal{O}_n$ and suppose that there exists $y_0 \in \overline{\mathcal{O}}_n$, $T > 0$, and an A_n-admissible path such that $\xi(0) = x_0$, $\xi(T) = y_0$. We have

$$|\dot{\xi}(t)|^4 \leq \langle A_n(\xi(t))\dot{\xi}(t), \dot{\xi}(t)\rangle$$

for a.e. $t \in [0, T]$, so

$$|\dot{\xi}(t)| \leq \|A_n^{\frac{1}{2}}(\xi(t))\| \qquad \text{for a.e. } t. \tag{4.13}$$

Since $A_n^{\frac{1}{2}}$ is smooth, we can select $L > 0$ verifying

$$\|A_n^{\frac{1}{2}}(x) - A_n^{\frac{1}{2}}(y)\| \leq L|x - y| \qquad \text{for any } x, y \in \overline{\mathcal{O}}_n. \tag{4.14}$$

So using (4.13) and (4.14), we get

$$|y - x_0| = |\xi(T) - \xi(0)| \leq L \int_0^T |\xi(t) - x_0| dt,$$

which implies $y_0 = x_0$ by Gronwall's inequality, concluding the proof of (i).

To establish (ii), fix n and, for any $x \in \mathcal{O}_n$, $b \in T_{\mathcal{O}_n}(x) = \mathbb{R}^N$, set

$$\|b\|_x = \langle g_n(x)b, b\rangle. \tag{4.15}$$

The previous formula gives the norm induced by the Riemannian metric (4.12) on $T_{\mathcal{O}_n}(x)$. We claim that

$$\|b\|_x \leq 1 \quad \text{if and only if} \quad b \otimes b \leq A_n(x). \tag{4.16}$$

Assume $\|b\|_x \leq 1$. By the Hölder inequality,

$$\langle A_n(x)a, b\rangle^2 \leq \langle A_n(x)a, a\rangle\langle A_n(x)b, b\rangle \qquad \text{for } a \in \mathbb{R}^N.$$

We deduce

$$\begin{aligned}\langle (b \otimes b - A_n(x))a, a\rangle &= \langle b, a\rangle^2 - \langle A_n(x)a, a\rangle \\ &\leq \|b\|_x\langle A_n(x)a, a\rangle - \langle A_n(x)a, a\rangle \leq 0.\end{aligned}$$

Conversely, let $b \otimes b \leq A_n(x)$. Then

$$0 \geq \langle (b \otimes b - A_n(x))a, a\rangle = \|b\|_x^2 - \|b\|_x,$$

which completes the proof of the claim. Now let us name g_n-admissible any continuous piecewise C^1 curve contained in $\mathcal{O}_n$ which satisfies

$$\|\dot\xi(t)\|_{\xi(t)} \leq 1 \qquad \text{for a.e. } t$$

and recall that the Riemannian distance between $x, y \in \mathcal{O}_n$ is given by

$$\begin{aligned}d_{g_n}(x, y) = \inf\{T &: \text{there exists } \xi \\ &g_n\text{-admissible with } \xi(0) = x,\ \xi(T) = y\}.\end{aligned} \tag{4.17}$$

We observe that an A_n-admissible curve which joins two points of $\mathcal{O}_n$ is contained in $\mathcal{O}_n$ thanks to (i). Hence (4.16) permits us to assert that any curve joining points in $\mathcal{O}_n$ is A_n-admissible if and only if it is g_n-admissible. This together with (3.2) and (4.17) ends the proof. □

Remark 4.1. It is easily seen that the distances d_n are not topologically equivalent to the Euclidean distance. Nevertheless, by Proposition 4.2(ii) this equivalence holds restricted to $\mathcal{O}_n$. Hence, for any $x_0 \in \mathcal{O}_n$, the function $x \to d(x, x_0)$ is continuous in $\mathcal{O}_n$ (with respect to the Euclidean topology). We can also immediately verify that $\mathbb{R}^N$, with the metric given by d_n, is complete and that for any x, r, the balls $B_n(x, r)$ are relatively compact (in the Euclidean topology).

From now on we will consider only the Euclidean topology on $\mathbb{R}^N$. We set, for any $(x, t) \in \mathbb{R}^N \times [0, +\infty)$,

$$w_n(x, t) = \sup_{B_n(x,t)}\{u_0\} = \max_{\overline{B_n(x,t)}}\{u_0\}. \tag{4.18}$$

Proposition 4.3.

(i) *For any* $x \in \mathbb{R}^N$, $w_n(x, 0) = u_0(x)$.

(ii) $w_n \in X$.

Proof. Part (i) of the assertion follows directly from (4.18). To prove (ii), fix n and observe that, by Proposition 4.2(i) and (4.18), it follows that $w_n(x, t) = u_0(x)$ for any $(x, t) \in (\mathbb{R}^N \setminus \overline{\mathcal{O}}_n) \times [0, +\infty)$. Thus w_n is u.s.c. in this set because u_0 is u.s.c. Now let $(x, t) \in \overline{\mathcal{O}}_n \times [0, +\infty)$ and (x_k, t_k) be any sequence converging to this point. Then for every k, we can select y_k satisfying

$$y_k \in \overline{B_n(x_k, t_k)}, \qquad u_0(y_k) = w_n(x_k, t_k).$$

We claim that if $x \in \partial\mathcal{O}_n$, then y_k converges to x. In view of Proposition 4.2(i), we can suppose $\{y_k\}$, $\{x_k\} \subset \mathcal{O}_n$. Let $\sigma > 0$ and k be so large that $t_k < t + \sigma$.

Consider an A_n-admissible path ξ_k and $s_k \le t+\sigma$ such that $\xi_k(0) = x_k$, $\xi_k(s_k) = y_k$. Employing (4.13), (4.14) we get for any $t \in [0, s_k]$,

$$|\dot{\xi}_k(t)| \le \|A_n^{\frac{1}{2}}(\xi_k(t))\| \le L\|\xi_k(t) - x_k\| + \|A_n^{\frac{1}{2}}(x_k)\|.$$

Hence

$$|y_k - x_k| \le L \int_0^{s_k} |\xi_k(t) - x_k| dt + L(t_0 + \sigma)\|A_n^{\frac{1}{2}}(x_k)\|,$$

and by Gronwall's inequality we obtain

$$|y_k - x_k| \le L(t_0 + \sigma)\|A_n^{\frac{1}{2}}(x_k)\| \exp(L(t_0 + \sigma)).$$

So taking into account that x_k and $\|A_n^{\frac{1}{2}}(x_k)\|$ converge to x and 0, respectively, the claim follows. Therefore w_n is u.s.c. for $(x,t) \in \partial\mathcal{O}_n \times [0,+\infty)$.

Finally, let $x \in \mathcal{O}_n$. If n is sufficiently large, we have that $y_k \in \mathcal{O}_n$ and

$$d_n(y_k, x_0) \le t_k + d_n(y_k, x) \le t + \sigma \tag{4.19}$$

for any $\sigma > 0$. We can assume without loss of generality that y_k converge to a point $y_0 \in \mathcal{O}_n$ so that using (4.19) and the continuity of $y \to d_n(y,x)$, we find

$$d_n(x, y_0) = \lim_k d_n(x, y_k) \le t.$$

Hence $w_n(x,t) \ge u_0(y_0) \ge \limsup_k u_0(y_k) = \limsup_k w_n(x_k, t_k)$, which also gives the assertion in this case. □

5. Truncated problems.

We set, for any n, $(x,p) \in \mathbb{R}^N \times \mathbb{R}^N$,

$$H_n(x,p) = -\langle A_n(x)p, p\rangle^{\frac{1}{2}} \tag{5.1}$$

and we consider the equation

$$u_t(x,t) + H_n(x, \nabla u(x,t)) = 0. \tag{5.2}$$

Our aim is to show that the function w_n defined in (4.18) is a solution of (5.2), (2.2) in the sense of Definition 2.2.

We start by recalling some elementary facts from Riemannian geometry that we will apply to the Riemannian manifolds $(\mathcal{O}_n, g_n)$. In the next proposition $(\mathcal{M}, g)$ is a Riemannian manifold, d_g is the distance associated to g, and $B_g(x,r) = \{y : d_g(x,y) < r\}$ for $x \in \mathcal{M}$, $r > 0$. By a minimizing geodesic we mean a curve $\xi : [0,T] \to \mathcal{M}$ satisfying $d_g(\xi(t), \xi(s)) = |t - s|$ for every $t, s \in [0,T]$. For such a geodesic, the Riemannian norm of $\dot{\xi}(t)$ equals 1 for any t.

PROPOSITION 5.1. *Let $x_0 \in \mathcal{M}$. There exists $\delta_0 = \delta(x_0) > 0$ such that if $\delta \in (0, \delta_0)$, then*

(i) *any two points of $B_g(x_0, 3\delta)$ can be joined by a unique minimizing geodesic.*

(ii) *d_g^2 is differentiable in $B_g(x_0, 3\delta) \times B_g(x_0, 3\delta)$.*

(iii) *for any $y_0 \in \overline{B_g(x_0, \delta)}$ and $z_0 \in \overline{B_g(y_0, \delta)}$, $z_0 \ne y_0$, denoting by ξ the minimizing geodesic joining y_0 to z_0 with $\xi(0) = y_0$, $\xi(T) = z_0$, we have that $\dot{\xi}(T)$ is equal to the Riemannian gradient at z_0 of the function $z \to d_g(y_0, z)$.*

Proof. See [16, vol. 1], and [7] for verification of this proposition. □

Part (iii) of the previous proposition is a version of the classical Gauss lemma. Recall that, as proved in Proposition 4.2(ii), one has $d_n = d_{g_n}$ in $\mathcal{O}_n$.

COROLLARY 5.1. *Let* $x \in \mathcal{O}_n$, $\delta_0 > 0$ *be as in Proposition* 5.1 *with* $(\mathcal{M}, g) = (\mathcal{O}_n, g_n)$, $y_0 \in \overline{B_n(x_0, \delta)}$ *with* $0 < \delta < \delta_0$. *If* $f : \mathcal{O}_n \to \mathbb{R}$ *is a differentiable function which attains a local minimum on* $\overline{B_n(y_0, \delta)}$ *at a point* $z_0 \neq y_0$, *then*

$$\nabla f(z_0)\dot{\xi}(T) = H_n(z_0, \nabla f(z_0)), \tag{5.3}$$

where ξ *is the minimal geodesic joining* y_0 *to* z_0 *with* $\xi(0) = y_0$, $\xi(T) = z_0$.

Proof. For any $y \in \mathcal{O}_n$, denote by $\langle \cdot, \cdot \rangle_y$ and $\| \cdot \|_y$, respectively, the inner product and the norm induced by g_n on $T_{\mathcal{O}_n}(y)$. We claim that

$$\dot{\xi}(T) = A_n(z_0)\nabla_2 d_n(y_0, z_0), \tag{5.4}$$

where $\nabla_2 d_n(y_0, z_0)$ is the Euclidean gradient of the function $z \to d_n(y_0, z)$ at z_0 and this function is differentiable by virtue of Proposition 5.1(ii). To prove the claim we observe that

$$\langle A_n(z_0)\nabla_2 d_n(y_0, z_0), b\rangle_z = \langle \nabla_2 d_n(y_0, z_0), b\rangle$$

for any $b \in \mathbb{R}^N$. This equality is equivalent to $A_n(z_0)\nabla_2 d_n(y_0, z_0)$ being the Riemannian gradient of $z \to d_n(y_0, z)$ at z_0. Hence we conclude by using Proposition 5.1(iii).

Since z_0 is a local minimum of f on $\overline{B_n(y_0, \delta)}$ we know that

$$\nabla f(z_0) = \lambda \nabla_2 d_n(y_0, z_0) \qquad \text{for some } \lambda \leq 0.$$

Therefore, noting that by (5.4)

$$1 = \|\dot{\xi}(T)\|^2_{z_0} = \langle A_n(z_0)\nabla_2 d_n(y_0, z_0), \nabla_2 d_n(y_0, z_0)\rangle,$$

we find

$$\begin{aligned}\langle \nabla f(z_0), \dot{\xi}(T)\rangle &= \lambda \langle A_n(z_0)\nabla_2 d_n(y_0, z_0), \nabla_2 d_n(y_0, z_0)\rangle^{\frac{1}{2}} \\ &= -\langle A_n(z_0)\lambda\nabla_2 d_n(y_0, z_0), \lambda\nabla_2 d_n(y_0, z_0)\rangle^{\frac{1}{2}} = H_n\left(z_0, \nabla f(z_0)\right). \qquad \square\end{aligned}$$

LEMMA 5.1. w_n *is a solution of equation* (5.2).

Proof. Since $w_n(x_0, t_0) \wedge M = \max_{\overline{B_n(x_0, t_0)}}\{u_0 \wedge M\}$ for any $(x_0, t_0) \in \mathbb{R}^N \times (0, +\infty)$, $M \in \mathbb{R}$, we need to consider only the case where w_n is finite and show that for any ψ, C^1 supertangent to w_n at (x_0, t_0), we have

$$\psi_t(x_0, t_0) + H_n(x_0, \nabla\psi(x_0, t_0)) = 0. \tag{5.5}$$

If $x_0 \in \mathbb{R}^N \setminus \mathcal{O}_n$, then $H_n(x_0, \nabla\psi(x_0, t_0)) = 0$ and the function $t \to w_n(x_0, t)$ is identically equal to $u_0(x_0)$, which gives $\psi_t(x_0, t_0) = 0$ and, consequently, (5.5).

Now we claim that for any $(x, t) \in \mathcal{O}_n \times (0, +\infty)$, $\sigma \in (0, t)$,

$$w_n(x, t) = \max_{\overline{B_n(x,\sigma)}} \{w_n(\cdot, t - \sigma)\}. \tag{5.6}$$

In fact since $\overline{B_n(y, t - \sigma)} \subset \overline{B_n(x, t)}$ for every $y \in B_n(x, \sigma)$, we see that

$$w_n(x, t) \geq \max_{\overline{B_n(x,\sigma)}} \{w_n(\cdot, t - \sigma)\}. \tag{5.7}$$

Moreover we can find z_0, z_1 verifying $u_0(z_1) = w_n(x,t)$, $d(z_0,x) \leq \sigma$, $d(z_0,z_1) \leq t-\sigma$ from which we deduce

$$w_n(x,t) = w_n(z_0, t-\sigma); \tag{5.8}$$

this equality together with (5.7) yields (5.6).

Fix $(x_0,t_0) \in \mathcal{O}_n \times (0,+\infty)$, $\delta \in (0, t_0 \wedge \delta_0)$, where δ_0 is as in Proposition 5.1 applied to $(\mathcal{O}_n, g_n)$ and select, using (5.6), $y_0 \in \overline{B_n(x,\delta)}$ such that

$$w_n(x_0,t_0) = w_n(y_0, t_0-\delta). \tag{5.9}$$

We can always assume that the C^1 supertangent ψ verifies $\psi(x_0,t_0) = w_n(x_0,t_0)$ and $\psi(x,t) \geq w_n(x,t)$ in $B_n(x_0, 3\delta) \times (t_0-\delta, t_0+\delta)$. We have, for any $x \in \overline{B_n(y_0,\delta)}$,

$$\begin{aligned}\psi(x,t_0) \geq w_n(x,t_0) &= \max_{\overline{B_n(x,\delta)}} \{w_n(\cdot, t_0-\delta)\} \\ &\geq w_n(y_0,t_0-\delta) = w_n(x_0,t_0) = \psi(x_0,t_0)\end{aligned}$$

so that

$$\psi(x_0,t_0) = \min_{\overline{B_n(y_0,\delta)}} \{\psi(\cdot, t_0)\}. \tag{5.10}$$

If $x_0 = y_0$, then by (5.10)

$$\nabla\psi(x_0,t_0) = 0. \tag{5.11}$$

Moreover because the function $t \to w_n(x_0,t)$ is nondecreasing by (4.18) and constant in $[t_0-\delta, t_0]$ by (5.9), we get

$$\psi_t(x_0,t_0) = 0. \tag{5.12}$$

This equality and (5.11) give (5.5).

If $x_0 \neq y_0$, let $d_n(x_0,y_0) = \rho \leq \delta$ and denote by ξ the minimizing geodesic joining y_0 to x_0 with $\xi(0) = y_0$ and $\xi(\rho) = x_0$. The path ξ can be extended to an interval $[0,\rho')$, $\rho' > \rho$, with the property

$$d_n(\xi(t), y_0) = t, \qquad t \in [0,\rho'). \tag{5.13}$$

Set $f(t) = \psi(\xi(t), t+t_0-\rho)$ and employ (5.6), (5.13), and (5.9) to find for t close to ρ,

$$\begin{aligned}f(t) \geq w_n(\xi(t), t+t_0-\delta) &= \max_{\overline{B_n(\xi(t),t)}} \{w_n(\cdot, t_0-\delta)\} \\ &\geq w_n(y_0,t_0-\delta) = w_n(x_0,t_0) = f(\rho).\end{aligned}$$

Hence

$$\frac{df}{dt}(\rho) = \nabla\psi(x_0,t_0)\cdot\dot{\xi}(\rho) + \psi_t(x_0,t_0) = 0.$$

By this equality, (5.10), and Corollary 5.1, we get (5.5) and conclude the proof. □

Now we show that the sequence w_n converges in a suitable sense to a function w defined via

$$w(x,t) = \inf_{\varepsilon>0} \sup_{B(x,t+\varepsilon)} \{u_0\}. \tag{5.14}$$

This result along with stability properties of viscosity solutions will imply that w is a solution of (2.1).

Definition 5.1. *Given a sequence ϕ_n of functions from $\mathbb{R}^N \times [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ and $\phi : \mathbb{R}^N \times [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$, we say that $\phi = \limsup^*_n \phi_n$ in $\mathbb{R}^N \times [0,+\infty)$ if for any $(z,t) \in \mathbb{R}^N \times [0,+\infty)$,*

$$\phi(z,t) = \sup\left\{\limsup_n \phi_n(z_n,t_n) : (z_n,t_n) \to (z,t)\right\}.$$

Proposition 5.2. *Let $\phi = \limsup^*_n \phi_n$ in $\mathbb{R}^N \times [0,+\infty)$. One has that*

(i) *ϕ is u.s.c.*

(ii) *$\phi \wedge M = \limsup^*_n(\phi_n \wedge M)$ in $\mathbb{R}^N \times [0,+\infty)$ for any $M \in \mathbb{R}$.*

(iii) *If $\overline{\phi} : \mathbb{R}^N \times [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ is u.s.c. and $\overline{\phi} \geq \phi_n$ on $\mathbb{R}^N \times [0,+\infty)$ for every $n \in \mathbb{N}$, then $\overline{\phi} \geq \phi$.*

Proof. For verification of Proposition 5.2, see [11], where the previous limit is called $\Gamma - \limsup$. □

Lemma 5.2. *$w = \limsup^*_n w_n$ in $\mathbb{R}^N \times [0,+\infty)$.*

Proof. The proof is based on the inclusions

$$B_n(x_n,t_n) \subset B(x_0,t_0+\delta), \tag{5.15}$$

$$B\left(x_0, t_0 + \frac{\delta}{2}\right) \subset B_n(x_0,(1+\delta)(t_0+\delta)), \tag{5.16}$$

which we claim are true for any $(x_0,t_0) \in \mathbb{R}^N \times [0,+\infty)$, (x_n,t_n) converging to this point, $\delta > 0$, and n sufficiently large.

To obtain the first one choose n such that $d(x_n,x_0) < \frac{\delta}{2}$, $t_n < t_0 + \frac{\delta}{2}$ and note that, by this last inequality, for any $y \in B_n(x_n,t_n)$, there exists an A_n-admissible path ξ with $\xi(0) = x_n$, $\xi(t_0 + \frac{\delta}{2}) = y$. Furthermore, ξ is A-admissible since $A_n \leq A$; therefore,

$$d(x_0,y) \leq d(x_0,x_n) + d(x_n,y) < t_0 + \delta,$$

which is (5.15).

Now choose $z \in B(x_0, t_0 + \frac{\delta}{2})$ and an A-admissible path ζ verifying $\zeta(0) = x_0$, $\zeta(t_0+\delta) = z$. Since this curve is continuous we find

$$\zeta([0,t_0+\delta]) \subset B^E(0,r) \tag{5.17}$$

for a suitable $r > 0$. Exploiting the fact that A_n converges to A uniformly on compact subsets, we also have

$$A(x) \leq (1+\delta)^2 A_n(x), \qquad x \in B^E(0,r) \tag{5.18}$$

for n sufficiently large. By (5.17) and (5.18) the path $\overline{\zeta}(s) = \zeta(\frac{s}{1+\delta})$ satisfies

$$\begin{aligned}\dot{\overline{\zeta}}(s) \otimes \dot{\overline{\zeta}}(s) &= \frac{1}{(1+\delta)^2}\left(\dot{\zeta}\left(\frac{s}{1+\delta}\right) \otimes \dot{\zeta}\frac{s}{1+\delta}\right) \\ &\leq \frac{1}{(1+\delta)^2} A\left(\zeta\left(\frac{s}{1+\delta}\right)\right) \leq A_n(\overline{\zeta}(s))\end{aligned}$$

for a.e. $s \in [0,(1+\delta)(t_0+\delta)]$ and, obviously, $\overline{\zeta}(0) = x_0$, $\overline{\zeta}((1+\delta)(t_0+\delta)) = z$. Therefore $d_n(x_0,z) \le (1+\delta)(t_0+\delta)$, so (5.16) follows.

The relation (5.15) implies

$$w_n(x_n,t_n) \le \sup_{B(x_0,t_0+\delta)} \{u_0\} \le w(x_0,t_0+\delta)$$

for any given $\delta > 0$ and n sufficiently large. Therefore, taking into account that w is u.s.c., we have

$$w(x_0,t_0) \ge \limsup_{\delta\to 0} w(x_0,t_0+\delta) \ge \limsup_n w_n(x_n,t_n).$$

To end the proof we construct a sequence (z_n,s_n) converging to (x_0,t_0) such that

$$\limsup_n w_n(z_n,s_n) \ge w(x_0,t_0). \tag{5.19}$$

Suppose first that $w(x_0,t_0) < +\infty$ and fix $\delta > 0$. For any $\sigma > 0$ we can select by (5.14) a point $z_0 \in B(x_0,t_0+\frac{\delta}{2})$ satisfying

$$u_0(z_0) \ge \sup_{B(x_0,t_0+\frac{\delta}{2})} \{u_0\} - \sigma \ge w(x_0,t_0) - \sigma.$$

From this and (5.16) we obtain

$$w_n(x_0,(1+\delta)(t_0+\delta)) \ge u_0(z_0) \ge w(x_0,t_0) - \sigma.$$

Hence

$$\liminf_n w_n(x_0,(1+\delta)(t_0+\delta)) \ge w(x_0,t_0). \tag{5.20}$$

Then we choose δ_h converging to zero and set $t_h = (1+\delta_h)(t_0+\delta_h)$. By (5.20), there exists a subsequence w_{k_h} of w_n such that

$$w_{k_h}(x_0,t_h) \ge w(x_0,t_0) - \frac{1}{h}. \tag{5.21}$$

If we define

$$\begin{aligned} z_n &= x_0 \qquad \text{for any } n,\\ s_n &= t_h \qquad \text{for } k_h \le n < k_{h+1}, \end{aligned}$$

from the fact that $w_{k_h}(z_{k_h},s_{k_h}) = w_{k_h}(x_0,t_h)$ is a subsequence of $w_n(z_n,s_n)$ and by using (5.21), we get (5.19).

In the case where $w(x_0,t_0) = +\infty$, it is possible to argue as above to discover that $\lim_n w_n(x_0,(1+\delta)(t_0+\delta)) = +\infty$ for every $\delta > 0$. So, denoting by δ_n any sequence converging to zero and setting $s_n = (1+\delta_n)(t_0+\delta_n)$, $z_n = x_0$, we find that (5.19) holds. □

Remark 5.1. It is worth noting that, as easily seen in Example 8.1, the sequence w_n may not in general converge pointwisely to w. If instead $(\mathbb{R}^N,d)$ is complete, then the convergence of w_n to w is uniform on compact sets of $\mathbb{R}^N$.

THEOREM 5.1. *w is a solution of* (2.1), (2.2).

Proof. $w \in X$ thanks to Proposition 5.2(i). To prove that w is a solution of equation (2.1), we only need to show that H_n converges to H uniformly on a compact subset of $\mathbb{R}^N \times \mathbb{R}^N$, since in this case the assertion is a consequence of Lemmata 5.1, 5.2 and well-known stability results for viscosity solutions (see [2]). To get the convergence of H_n, let $\mathcal{K}$ be a compact subset of $\mathbb{R}^N \times \mathbb{R}^N$ and $m > 0$ verifying $A \geq m^2 I$ in $\mathcal{K}_1 = \{x : (x,p) \in \mathcal{K} \text{ for some } p\}$ and $|p| \leq \frac{1}{m}$ in $\mathcal{K}_2 = \{p : (x,p) \in \mathcal{K} \text{ for some } x\}$. Then for any $(x,p) \in \mathcal{K}$, we have

$$|H_n(x,p) - H(x,p)| \leq \frac{1}{m^2} \max_{\mathcal{K}_1} \|A_n(x) - A(x)\|$$

and we conclude by employing the uniform convergence of A_n to A on $\mathcal{K}_1$.

We must still prove that

$$w(x,0) = u_0(x) \qquad \text{for every } x \in \mathbb{R}^N. \tag{5.22}$$

The inequality $w(x,0) \geq u_0(x)$ is clear from (5.14). For any fixed x, according to Proposition 3.1(ii), we have that $B(x,\varepsilon)$ is relatively compact for $\varepsilon > 0$ sufficiently small and so for such ε there exists $x_\varepsilon \in \overline{B(x,\varepsilon)}$ satisfying

$$u_0(x_\varepsilon) = \sup_{B(x,\varepsilon)} \{u_0\}.$$

Therefore, by (5.14), $w(x,0) = \lim_{\varepsilon\to 0} u_0(x_\varepsilon)$ and since $u_0 \in X_0$, we get $w(x,0) \leq u_0(x)$. The equality (5.22) is therefore proved. □

In what follows we will refer to w as the canonical solution of problem (2.1), (2.2).

We conclude this section pointing out an immediate consequence of Lemma 5.2 and of Proposition 5.2(iii) that we will use in the comparison results of section 6.

COROLLARY 5.2. *Let u be a solution of* (2.1), (2.2) *and assume that $u \geq w_n$ on $\mathbb{R}^N \times [0,+\infty)$ for every $n \in \mathbb{N}$. Then $u \geq w$.*

6. Construction of a test function.

We consider a solution u of (2.1) and for any $\alpha \in \mathbb{R}$, the level sets

$$\Gamma_\alpha(t) = \{x : u(x,t) \geq \alpha\}, \qquad t \in [0,+\infty),$$

which are closed since $u \in X$, possibly empty. We set

$$J_\alpha = \{t \geq 0 : \Gamma_\alpha(t) \neq \emptyset\}$$

and define on $\mathbb{R}^N \times J_\alpha$ the function

$$v_\alpha : (x,t) \to -d^E(x, \Gamma_\alpha(t)).$$

The purpose of this section is to show that v_α is locally Lipschitz and verifies with its generalized gradient a certain differential inequality related to equation (2.1). These properties will be crucial for the proof of the comparison results of the next section.

We fix $\alpha \in \mathbb{R}$ and we write in this section for simplicity Γ, J, v instead of Γ_α, J_α, v_α, respectively. We start with two lemmata.

LEMMA 6.1.

(i) *If J is not empty, then $J \setminus \{0\}$ is a open unbounded interval.*

(ii) *Let $\mathcal{H}$ be a compact set of $\mathbb{R}^N$ and define $J_{\mathcal{H}} = \{t > 0 : \Gamma(t) \cap \mathcal{H} \neq \emptyset\}$. If $J_{\mathcal{H}}$ is not empty, then it is a closed unbounded interval. Moreover if $\bar{t} = \min\{J_{\mathcal{H}}\}$ is positive, then $\Gamma(\bar{t}) \cap \mathcal{H} \subset \partial\mathcal{H}$.*

Proof. If J is not empty, then it is an unbounded interval by the monotonicity of u; see Remark 2.1. To show that $J \setminus \{0\}$ is open we argue by contradiction and assume that J has a minimum t_0 strictly positive. Then (x_0, t_0) is a minimum of the function $\psi : (x,t) \to t$ on $\{u \geq \alpha\}$ for every $x_0 \in \Gamma(t_0)$. This implies by Proposition 2.1 that equality (2.5) holds, which is impossible because of the definition of ψ.

Now we prove (ii). Clearly $J_{\mathcal{H}}$ is an unbounded interval. To show that it is closed, set $\bar{t} = \inf J_{\mathcal{H}}$ and denote by t_n a sequence of $J_{\mathcal{H}}$ converging to $\bar{t}$. Hence there exist $x_n \in \mathcal{H}$ which verify $u(x_n, t_n) \geq \alpha$. Moreover the sequence x_n converges, up to a subsequence, to a point $\overline{x} \in \mathcal{H}$. Since u is u.s.c. one has $u(\overline{x}, \bar{t}) \geq \alpha$, which gives $\bar{t} \in J_{\mathcal{H}}$. Finally if $\bar{t} > 0$ and $\Gamma(\bar{t}) \cap \mathcal{H}^\circ \neq \emptyset$, we repeat the argument of the first part to get a contradiction. This ends the proof. □

LEMMA 6.2.

(i) *v is u.s.c. in $\mathbb{R}^N \times J$.*

(ii) *For any $(x_0, t_0) \in \mathbb{R}^N \times J$, $t_0 > 0$, and ψ C^1-supertangent to v at (x_0, t_0), one has*

$$|\nabla \psi(x_0, t_0)| \leq 1. \tag{6.1}$$

Proof. Fix $(x_0, t_0) \in \mathbb{R}^N \times J$ and let (x_n, t_n) be a sequence in $\mathbb{R}^N \times J$ converging to (x_0, t_0) such that $\lim_n v(x_n, t_n)$ exists. This limit is then finite because by employing the monotonicity of u, we can choose $\varepsilon > 0$ verifying $(t_0 - \varepsilon) \in J$ and $|v(x_n, t_n)| \leq |v(x_0, t_0 - \varepsilon)| + \varepsilon$ for n sufficiently large. Because $\Gamma(t_n)$ is a closed nonempty set, there exists, for every n, $y_n \in \Gamma(t_n)$ satisfying

$$-|y_n - x_n| = v(x_n, t_n). \tag{6.2}$$

Hence the sequence y_n is bounded and so convergent, up to a subsequence, to a point y_0 which belongs to $\Gamma(t_0)$ by the upper semicontinuity of u. Therefore taking into account (6.2), we have

$$\lim_n v(x_n, t_n) = -|x_0 - y_0| \leq v(x, t),$$

and this proves (i).

The inequality (6.1) follows from the fact that $x \to v(x,t)$ is Lipschitz continuous with Lipschitz constant 1 for every t (see [1]). □

PROPOSITION 6.1.

(i) *v is locally Lipschitz in $\mathbb{R}^N \times J$.*

(ii) *For any compact set $\mathcal{K}$ of $\mathbb{R}^N \times (J \setminus \{0\})$, there exists a positive constant $L = L(\mathcal{K})$ such that*

$$s + H(x, p) \geq Lv(x, t) \tag{6.3}$$

for every $(x,t) \in \mathcal{K}$, $(p, s) \in \partial v(x,t)$.

Proof. As a first step let us show that for any $(x_0, t_0) \in \mathbb{R}^N \times J$, $t_0 > 0$, ψ C^1-supertangent to v at this point, one has

$$\psi_t(x_0, t_0) + H(z_0, \nabla \psi(x_0, t_0)) = 0, \tag{6.4}$$

where z_0 is any element of $\Gamma(t_0)$ satisfying

$$|z_0 - x_0| = -v(x_0, t_0). \tag{6.5}$$

In fact set $\beta = v(x_0,t_0)$, $b = x_0 - z_0$, $\overline{\psi}(x,t) = \psi(x+b,t)$ and note that ψ attains a local minimum at (x_0,t_0) on the set $\{v \geq \beta\}$. Then denote by U a neighborhood of (x_0,t_0) such that the inequality $\psi(x,t) \geq \psi(x_0,t_0)$ holds for every $(x,t) \in U \cap \{v \geq \beta\}$ and write $\widetilde{U} = \{(z,t) : (z+b,t) \in U\}$. If $(z,t) \in \widetilde{U} \cap \{u \geq \alpha\}$, then $(z+b,t) \in U \cap \{v \geq \beta\}$ since $|b| = -\beta$. So we obtain $\overline{\psi}(z,t) \geq \overline{\psi}(z_0,t_0)$ which shows that (z_0,t_0) is a local minimum point of $\overline{\psi}$ on $\{u \geq \alpha\}$. Therefore by Proposition 2.1 we find

$$\overline{\psi}_t(z_0,t_0) + H(z_0, \nabla\overline{\psi}(z_0,t_0)) = 0$$

and (6.4) follows.

We turn now to the proof of (i). Let $\mathcal{K}$ be a compact subset of $\mathbb{R}^N \times J$. We can assume without loss of generality that $\mathcal{K} = \mathcal{H} \times I$, where $\mathcal{H}$ and I are a compact domain of $\mathbb{R}^N$ and a compact interval contained in J, respectively. Set $t_1 = \min I$, $m = \min_{\mathcal{H}}\{v(\cdot,t_1)\}$. Since, by the monotonicity of v, $m = \min_{\mathcal{K}}\{v\}$, we have that if $(x_0,t_0) \in \mathcal{K}$ and z_0 verifies (6.5), then $z_0 \in \widetilde{\mathcal{H}} = \{x : d^E(x,\mathcal{H}) \leq -m\}$ which is a compact set. From this, (6.4), and Lemma 6.2(ii) we get for $t_0 > 0$,

$$|\psi_t(x_0,t_0)| \leq M \tag{6.6}$$

for any positive constant M satisfying

$$\|A(x)\| \leq M^2, \qquad x \in \widetilde{\mathcal{H}}.$$

Since the inequality (6.6) holds for any $(x_0,t_0) \in \mathcal{K}$, $t_0 > 0$, and any ψ C^1 supertangent to v at (x_0,t_0), we discover that (see [4], [10]) for every $x \in \mathcal{H}$, the function $t \to v(x,t)$ is Lipschitz continuous in I with Lipschitz constant less than or equal to M. Then we get the assertion (i) recalling that v is also Lipschitz continuous in x with Lipschitz constant independent of t.

To prove (ii) let $\mathcal{K}$ be a compact subset of $\mathbb{R}^n \times (J \setminus \{0\})$ and let $\mathcal{H}$, $\widetilde{\mathcal{H}}$, and I be as above. We first select two positive constants $\widetilde{L}$, $\widetilde{m}$ satisfying

$$\begin{gathered}\|A(x) - A(y)\| \leq \widetilde{L}|x-y|,\\ A(x) \geq \widetilde{m}^2 I\end{gathered}$$

for any $x, y \in \widetilde{\mathcal{H}}$. Then we compute for any $p \neq 0$, $x, y \in \widetilde{\mathcal{H}}$,

$$\begin{aligned}|H(x,p) - H(y,p)| &\leq \frac{\|A(x)-A(y)\||p|^2}{\langle A(x)p,p\rangle^{\frac{1}{2}} + \langle A(y)p,p\rangle^{\frac{1}{2}}}\\ &\leq \frac{\widetilde{L}}{\widetilde{m}}|x-y|\,|p|.\end{aligned} \tag{6.7}$$

Set $L = \widetilde{L}/\widetilde{m}$, recall Lemma 6.2, and use (6.7) in (6.4) to find for $(x_0,t_0) \in \mathcal{K}$,

$$\begin{aligned}&\psi_t(x_0,t_0) + H(x_0, \nabla\ psi(x_0,t_0))\\ &\quad \geq -L|\nabla\psi(x_0,t_0)||x_0 - z_0| = Lv(x_0,t_0).\end{aligned} \tag{6.8}$$

If, in particular, v is differentiable at (x_0,t_0), from (6.8) we get

$$v_t(x_0,t_0) + H(x_0, \nabla v(x_0,t_0)) \geq Lv(x_0,t_0). \tag{6.9}$$

Hence we obtain (6.3) by using Proposition 2.2(iv) and the fact that H is concave in p. □

7. Comparison results.

We start by proving a maximum principle for equation (2.1). Given two solutions u_1, u_2 of (2.1), we will put by convention $u_1(x,t) - u_2(x,t) = 0$ whenever u_1 and u_2 are both infinite at (x,t).

THEOREM 7.1. *Let $\mathcal{H}$ be a bounded set of $\mathbb{R}^N$ and I be any interval with minimum equal to* 0*. If u_1 and u_2 are solutions of* (2.1), (2.2), *then*

$$\sup_{\mathcal{H}\times I}\{|u_1 - u_2|\} = \sup_{\partial\mathcal{H}\times I}\{|u_1 - u_2|\}. \tag{7.1}$$

Proof. We argue by contradiction and assume that there exists $(x_0, t_0) \in \mathcal{H}^\circ \times I$ and two positive constants α, β satisfying

$$\alpha - \beta > \sup_{\partial\mathcal{H}\times I}\{|u_1 - u_2|\}, \tag{7.2}$$

$$u_1(x_0, t_0) \geq \alpha > \beta > u_2(x_0, t_0). \tag{7.3}$$

We define

$$\Gamma_1(t) = \{x : u_1(x,t) \geq \alpha\},$$
$$\Gamma_2(t) = \{x : u_2(x,t) \geq \beta\},$$

and

$$J_i = \{t \geq 0 : \Gamma_i(t) \cap \overline{\mathcal{H}} \neq \emptyset\}, \qquad i = 1, 2.$$

By Lemma 6.1, J_1 has a minimum that we denote by t_1. We wish to show that

$$[t_1, t_0] \subset J_2. \tag{7.4}$$

In fact if $t_1 = 0$, then for any $x \in \Gamma_1(t_1) \cap \overline{\mathcal{H}}$, one has

$$u_2(x, t_1) = u_1(x, t_1) \geq \alpha > \beta$$

which gives $t_1 \in J_2$ and so (7.4).

If $t_1 > 0$, then by Lemma 6.1(ii), $\Gamma_1(t_1) \cap \overline{\mathcal{H}} \subset \partial\mathcal{H}$. This implies

$$\Gamma_1(t_1) \cap \overline{\mathcal{H}} \subset \Gamma_2(t_1); \tag{7.5}$$

otherwise, for any $x \in (\Gamma_1(t_1) \cap \overline{\mathcal{H}}) \setminus (\Gamma_2(t_1))$, we should have

$$u_1(x,t) - u_2(x,t) > \alpha - \beta$$

which is impossible by (7.2). From (7.5) we immediately get (7.4). Now set $\mathcal{K} = \overline{\mathcal{H}} \times [t_1, t_0]$ and for any $(x,t) \in \mathcal{K}$,

$$v(x,t) = -d^E(x, \Gamma_2(t)).$$

By (7.4) we can use Proposition 6.1 to deduce that v is Lipschitz in $\mathcal{K}$. We fix $\lambda > L$, where L is as in formula (6.3) and define $\overline{v}(x,t) = e^{-\lambda t} v(x,t)$ for any $(x,t) \in \mathcal{K}$. From (6.3) and the positive homogeneity of H in p, we get

$$s + H(x,p) \geq (L - \lambda)\overline{v}(x,t) \geq 0 \tag{7.6}$$

for any $(x,t) \in \mathcal{K}$, $t > 0$, $(p,s) \in \partial\overline{v}(x,t)$. Now choose $(\overline{x},\overline{t}) \in \{u_1 \geq \alpha\} \cap \mathcal{K}$ such that

$$\overline{v}(\overline{x},\overline{t}) = \min_{\{u_1 \geq \alpha\}\cap\mathcal{K}} \{\overline{v}\}. \tag{7.7}$$

Since, by the definition of $\overline{v}$ and (7.3), $\overline{v}(x_0,t_0) < 0$ and $\overline{v}(\overline{x},\overline{t}) \leq \overline{v}(x_0,t_0)$, we find

$$\overline{v}(\overline{x},\overline{t}) < 0. \tag{7.8}$$

Keeping in mind the definition of $\overline{v}$, from (7.8) we deduce

$$u_2(\overline{x},\overline{t}) < \beta$$

and, consequently,

$$\overline{x} \in \mathcal{H}^\circ, \tag{7.9}$$

because otherwise, from the inequality $u_1(\overline{x},\overline{t}) - u_2(\overline{x},\overline{t}) > \alpha - \beta$, we should get a contradiction of (7.2).

We claim that

$$\overline{t} > t_1. \tag{7.10}$$

In fact if $t_1 = 0$, $t_1 = \overline{t}$ is impossible since this should imply $u_1(\overline{x},\overline{t}) = u_2(\overline{x},\overline{t})$. If instead $t_1 > 0$, then the equality $\overline{t} = t_1$ contradicts (7.9) by Lemma 6.1(ii). Thanks to (7.7), (7.9), and (7.10) we see that $(\overline{x},\overline{t})$ is a point of local minimum of $\overline{v}$ on $\{u_1 \geq \alpha\} \cap \mathbb{R}^N \times (t_1,t_0]$. Hence by Proposition 2.4 and Remark 2.2, we get the existence of $(\overline{p},\overline{s}) \in \partial\overline{v}(\overline{x},\overline{t})$ verifying

$$\overline{s} + H(\overline{x},\overline{p}) \leq 0.$$

This contradicts (7.6) if we take into account (7.8) and the inequality $\lambda > L$. □

The argument of the previous proof gives the following corollary.

COROLLARY 7.1. *Let u_1, u_2 be two solutions of* (2.1) *and I be an interval with minimum equal to* 0. *If $u_1(x,0) \leq u_2(x,0)$ for every $x \in \mathbb{R}^N$, then either*

$$u_1 \leq u_2 \qquad \text{in } \mathbb{R}^N \times I$$

or

$$\liminf_{R\to+\infty} \sup_{\partial B^E(0,R)\times I} \{u_1 - u_2\} > 0.$$

Proof. If there exists $(x_0,t_0) \in \mathbb{R}^N \times I$ such that $u_1(x_0,t_0) > u_2(x_0,t_0)$, use the argument of Theorem 7.1 to show that

$$\sup_{\partial B^E(0,R)\times I} \{u_1 - u_2\} \geq u_1(x_0,t_0) - u_2(x_0,t_0)$$

if $R \geq |x_0|$. This proves the assertion. □

The next result yields the minimality of the canonical solution among the solutions of (2.1), (2.2).

THEOREM 7.2. *The canonical solution w is the minimal solution of* (2.1), (2.2).

Proof. In view of Corollary 5.2 it is enough to prove that if u is a solution of (2.1), (2.2), then for any n,

$$u \geq w_n \qquad \text{in } \mathbb{R}^N \times (0, +\infty), \tag{7.11}$$

where w_n are the functions defined in (4.18). Note that w_n is a subsolution of (2.1), (2.2) by (4.2). So if (7.11) is not true, using Corollary 7.1 one has that there exists $\overline{n}$ such that

$$\liminf_{R \to +\infty} \sup_{\partial B^E(0,R) \times [0,+\infty)} \{w_{\overline{n}} - u\} > 0. \tag{7.12}$$

For R sufficiently large, $\partial B^E(0,R) \subset \mathbb{R}^N \setminus \mathcal{O}_{\overline{n}}$ and $w_{\overline{n}}(x,t) = u_0(x)$ for every $(x,t) \in (\mathbb{R}^N \setminus \mathcal{O}_{\overline{n}}) \times [0, +\infty)$ by virtue of Proposition 4.2(i) and (4.18). This contradicts (7.12) since u is nondecreasing. □

Using Corollary 7.1 and Theorem 7.2 we establish that problem (2.1), (2.2) is uniquely solvable when the solutions are required to satisfy a certain growth condition at infinity.

Corollary 7.2. *Let I be any interval with minimum equal to* 0 *and u be a solution of* (2.1), (2.2). *If*

$$\liminf_{R \to +\infty} \sup_{\partial B^E(0,R) \times I} (u - w) \leq 0, \tag{7.13}$$

then $u = w$ in $\mathbb{R}^N \times I$.

We now give a characterization for solutions of equation (2.1).

Theorem 7.3. *A function $u \in X$ is a solution of* (2.1) *if and only if one has*

$$u(x_0, t_0 + s) = \max_{\overline{B(x_0,s)}} \{u(\cdot, t_0)\} \tag{7.14}$$

for any $(x_0, t_0) \in \mathbb{R}^N \times [0, +\infty)$, $s < d(x_0, \infty)$.

Proof. Let x_0, t_0, s, and u be as in the statement above and denote by $\widetilde{w}$ the canonical solution of (2.1) with initial datum $x \to u(x, t_0)$, i.e., the function given by the formula (4.18) with u_0 replaced by $u(\cdot, t_0)$. We claim that

$$\widetilde{w}(x_0, s) = \max_{\overline{B(x_0,s)}} \{u(\cdot, t_0)\}. \tag{7.15}$$

In fact from the definition of $\widetilde{w}$, the upper semicontinuity of u_0, and the compactness of $\overline{B(x_0, s + \frac{1}{n})}$ for n sufficiently large, we see that

$$\widetilde{w}(x_0, s) = \lim_n \max_{\overline{B(x_0,s+\frac{1}{n})}} \{u(\cdot, t_0)\} \geq \max_{\overline{B(x_0,s)}} \{u(\cdot, t_0)\}. \tag{7.16}$$

We select x_n verifying

$$\begin{aligned} x_n &\in \overline{B\left(x_0, s + \frac{1}{n}\right)}, \\ u(x_n, t_0) &= \max_{\overline{B(x_0,s+\frac{1}{n})}} \{u(\cdot, t_0)\}. \end{aligned} \tag{7.17}$$

If $\widetilde{x}$ is a limit point of x_n, we have

$$\widetilde{x} \in \overline{B(x_0, s)}.$$

Taking (7.16) into account,

$$\max_{\overline{B(x_0,s)}}\{u(\cdot,t_0)\} \geq u(\widetilde{x},t_0) \geq \widetilde{w}(x_0,s) \geq \max_{\overline{B(x_0,s)}}\{u(\cdot,t_0)\}. \tag{7.18}$$

The claim implies that a function $u \in X$ satisfying (7.14) is solution of (2.1).

Conversely let u be a solution of (2.1). According to (7.15), equality (7.14) is equivalent to

$$u(x_0,t_0+s) = \widetilde{w}(x_0,s) \qquad \text{for } s < d(x_0,\infty). \tag{7.19}$$

If (7.19) is not true, by Theorem 7.2 there exists $\widetilde{t} \in (0, d(x_0,\infty))$ such that

$$u(x_0,t_0+\widetilde{t}) > \widetilde{w}(x_0,\widetilde{t}). \tag{7.20}$$

We define a function $\overline{u}_0 \in X_0$ via

$$\overline{u}_0 = \begin{cases} u(\cdot,t_0) & \text{in } B(x_0,\overline{t}), \\ u(\cdot,t_0) \vee \alpha & \text{in } \mathbb{R}^N \setminus B(x_0,\overline{t}), \end{cases} \tag{7.21}$$

where $\alpha = u(x_0,t_0+\widetilde{t})$ and $\widetilde{t} < \overline{t} < d(x_0,\infty)$ and denote by $\overline{w}$ the canonical solution of (2.1) with initial datum $\overline{u}_0$.

Now we observe that by (7.21),

$$u(x,t_0) \leq \overline{u}_0(x) = \overline{w}(x,0) \qquad \text{for every } x \tag{7.22}$$

and using (7.15), (7.20), and (7.21) we see that

$$\overline{w}(x_0,\widetilde{t}) = \widetilde{w}(x_0,\widetilde{t}) < u(x_0,t_0+\widetilde{t}) = \alpha. \tag{7.23}$$

Taking (7.22), (7.23) into account and applying Corollary 7.1 to $\overline{u}(x,t) = u(x,t_0+t) \wedge \alpha$ and $\overline{w}$, which are solutions of (2.1), we get

$$\liminf_{R\to+\infty} \sup_{\partial B^E(0,R)\times[0,+\infty)} \{\overline{u} - \overline{w}\} > 0.$$

This is impossible since

$$\overline{u} \leq \alpha \qquad \text{in } \mathbb{R}^N \times [0,+\infty)$$

and

$$\overline{w}(x,t) \geq \overline{u}_0(x) \geq \alpha \qquad \text{for any } (x,t) \in (\mathbb{R}^N \setminus B(x_0,\overline{t})) \times [0,+\infty). \qquad \square$$

The previous theorem has the following interpretation: As long as the ball $B(x_0,s)$ is relatively compact, every solution of (2.1) in the interval $[t_0,t_0+s]$ can be represented with a formula analogous to that of the canonical solution.

Remark 7.1. From Theorem 7.3 it easily follows that if the initial data satisfy

$$\lim_{|x|\to+\infty} u_0(x) = +\infty,$$

then the solution of (2.1),(2.2) is unique.

Moreover if $(\mathbb{R}^N, d)$ is complete, then every solution starting from a continuous u_0 is continuous in $\mathbb{R}^N \times [0,+\infty)$.

In the final part of the section we will prove that the uniqueness of solutions of (2.1), (2.2) for any initial datum is the equivalent of the metric space $(\mathbb{R}^N, d)$ being complete.

PROPOSITION 7.1. *If* $(\mathbb{R}^N, d)$ *is not complete, then*

$$u(x,t) = t - d(x,\infty), \qquad (x,t) \in \mathbb{R}^N \times (0,+\infty) \tag{7.24}$$

is a noncanonical solution of equation (2.1).

Proof. We have for any $x_0, y_0 \in \mathbb{R}^N$,

$$d(x_0,\infty) \leq d(x_0,y_0) + d(y_0,\infty), \tag{7.25}$$

so u is locally Lipschitz with respect to the Euclidean distance. By (7.25) we get

$$u(x_0, t+t_0) = -d(x_0,\infty) + t + t_0 \geq -d(y,\infty) + t_0$$

for any $t_0, t > 0$, $y \in \overline{B(x_0,t)}$. By Lemma 3.2 we know that if $t < d(x_0,\infty)$, there exists $z \in \partial B(x_0,t)$ such that

$$u(x_0, t+t_0) = -d(z,\infty) + t_0.$$

Therefore u satisfies (7.14) and is a solution of (2.1) according to Theorem 7.3. □

From the previous proposition and Theorem 7.3 we obtain the following corollary.

COROLLARY 7.3. *Problem* (2.1), (2.2) *is uniquely solvable for any initial datum if and only if the metric space* $(\mathbb{R}^N, d)$ *is complete.*

8. Examples. The first two examples illustrate the pathologies which can occur in problem (2.1), (2.2). As we explained before these phenomena depend on the lack of completeness of the distance associated to A.

The first example shows that the canonical solution can be discontinuous and infinite even starting from a smooth initial datum.

Example 8.1. Set $N = 2$ and put in (2.1), (2.2),

$$A(x_1,x_2) = \begin{bmatrix} x_1^4 + 1 & -x_1^2 \\ -x_1^2 & 1 \end{bmatrix}$$

and

$$u_0(x_1,x_2) = -x_1^2 x_2.$$

As usual let us denote by w the canonical solution and by x_0 the point $(1,1)$. We have the following assertions.

PROPOSITION 8.1.

(i) $w(x_0, 1-\delta) \leq 0$ *for every* $1 > \delta > 0$.

(ii) $w(x_0, 1) = +\infty$.

Proof. Consider any A-admissible path $\xi(t) = (\xi_1(t), \xi_2(t))$, which, by definition, verifies

$$\langle (\dot\xi(t) \otimes \dot\xi(t)) b, b\rangle \leq \langle A(\xi(t)) b, b\rangle \quad \text{a.e.} \tag{8.1}$$

for any $b \in \mathbb{R}^2$. Then put $b = (0, \dot\xi_2(t))$ in (8.1) to find

$$|\dot\xi_2(t)| \leq 1 \qquad \text{a.e.} \tag{8.2}$$

The last inequality implies in particular that

$$B(x_0, 1-\delta) \subset \{x_2 \geq 0\} \qquad \text{for every } 1 > \delta > 0,$$

which is equivalent to assertion (i).

Now fix $\varepsilon > 0$ and let $\sigma = \sigma(\varepsilon) > 0$ be such that $d(x_0, \widetilde{x}) < \varepsilon$ if $\widetilde{x} = (1, 1-\sigma)$. The path $\xi(t) = (\frac{1}{1-t}, 1-\sigma-t)$ is A-admissible as it can be seen by a direct calculation. Therefore $\xi(t) \in B(\widetilde{x}_0, 1) \subset B(x_0, 1+\varepsilon)$ for $t \in [0,1)$ and letting $t \to 1$, we discover that

$$\sup_{B(x_0, 1+\varepsilon)} \{u_0\} = +\infty.$$

Since $\varepsilon > 0$ is arbitrary, this gives (ii). □

The second example concerns the lack of uniqueness for solutions of (2.1), (2.2).

Example 8.2. Set $N = 1$, $u_0 \equiv 0$, and

$$A(x) = \begin{cases} 1 & \text{if } x \leq 1, \\ x^6 & \text{if } x \geq 1. \end{cases}$$

The distance d associated to A coincides with the Euclidean distance in $(-\infty, 1)$. If $x, y \in [1, +\infty)$ and $y \geq x$, then $\zeta(t) = \sqrt{\frac{x^2}{1-2tx^2}}$, $t \in [0, \frac{|x^2-y^2|}{2x^2y^2}]$ is A-admissible and it is the minimal path joining x to y. Therefore we have

$$d(x,y) = \begin{cases} |x-y| & \text{if } x, y \in (-\infty, 1), \\ \frac{|x^2-y^2|}{2x^2y^2} & \text{if } x, y \in [1, +\infty), \\ 1 - x + \frac{y^2-1}{2y^2} & \text{if } x \in (-\infty, 1),\ y \in [1, +\infty). \end{cases}$$

The number $d(x, \infty)$ defined in section 2 is given by

$$d(x, \infty) = \begin{cases} \frac{3}{2} - x, & x \in (-\infty, 1), \\ \frac{1}{2x^2}, & x \in [1, +\infty). \end{cases}$$

A family u_α, $\alpha \geq 0$, of solutions of (2.1), (2.2) with $u_0 \equiv 0$ is defined via (see Proposition 7.1)

$$u_\alpha(x,t) = \max\{\alpha t - \alpha d(x, \infty), 0\}.$$

For $\alpha > 0$ these solutions are not canonical.

Finally we show that even if the Hamiltonian H has superlinear growth in x, then the distance induced can be complete. That is, the converse of Corollary 3.1 is not true.

Example 8.3. Set $N = 2$,

$$A(x_1, x_2) = \begin{bmatrix} 1 & 0 \\ 0 & f(x_1) \end{bmatrix},$$

where f is any strictly positive locally Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$. If ξ is an A-admissible path starting from $x_0 = (x_1^0, x_2^0)$, it is easy to see that

$$|\dot{\xi}(t)|^2 \leq \max\{f(x_1) : x_1 \in [x_1^0 + t, x_1^0 - t]\} + 1 \qquad \text{a.e.}$$

Hence for any radius r, there exists a positive constant $M(x_0, r)$ such that

$$B(x_0, r) \subset B^E(x_0, M(x_0, r)r).$$

This inclusion shows that $B(x_0, r)$ is relatively compact and, consequently (see Proposition 3.1(iii)), the distance d is complete.

REFERENCES

[1] M. Bardi and I. Capuzzo Dolcetta, *Viscosity Solutions of Bellman Equation and Optimal Deterministic Control Theory*, Birkhäuser, Boston, to appear.

[2] G. Barles, *Solution de viscosité des equations de Hamilton-Jacobi*, Springer-Verlag, Berlin, 1994.

[3] G. Barles, *Discontinuous viscosity solutions of first order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.

[4] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 49–79.

[5] 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.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, PA, 1983.

[7] J. Cheeger and D. Ebin, *Comparison theorems in Riemannian geometry*, North-Holland, Amsterdam, 1975.

[8] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[9] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[10] M. G. Crandall and P. L. Lions, *Remarks on the existence and uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Illinois J. Math., 31 (1987), pp. 665–688.

[11] G. Dal Maso, *An introduction to Γ-convergence*, Birkhäuser, Boston, 1993.

[12] I. Ekeland and J. M. Lasry, *On the number of periodic trajectories for an Hamiltonian flow on a convex energy surface*, Ann. of Math. (2), 112 (1980), pp. 283–319.

[13] L. C. Evans and P. Souganidis, *Differential games and representation formulas for Hamilton-Jacobi equations*, Indiana U. Math. J., 33 (1984), pp. 773–795.

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

[15] M. Gromov, J. Lafontaine, and P. Pansu, *Structures métriques pour les variétés Riemanniennes*, CEDIC-FERNAND NATHAN, Paris, 1981.

[16] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, J. Wiley, New York, 1963.

[17] S. Koike, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with a degenerate coefficient*, Differential Integral Equations, 10 (1997), pp. 455–472.

[18] J. M. Lasry and P. L. Lions, *A remark on regularization in Hilbert space*, Israel J. Math., 55 (1986), pp. 257–266.

[19] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.

[20] A. Siconolfi, *A first order Hamilton–Jacobi equation with singularity and the evolution of level sets*, Comm. Partial Differential Equations, 20 (1995), pp. 277–308.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1446–1456, November 1997

STIFF SYSTEMS OF HYPERBOLIC CONSERVATION LAWS: CONVERGENCE AND ERROR ESTIMATES*

A. KURGANOV† AND E. TADMOR†‡

Abstract. We are concerned with 2×2 nonlinear relaxation systems of conservation laws of the form $u_t + f(u)_x = -\frac{1}{\delta}S(u,v)$, $v_t = \frac{1}{\delta}S(u,v)$ which are coupled through the stiff source term $\frac{1}{\delta}S(u,v)$. Such systems arise as prototype models for combustion, adsorption, etc. Here we study the convergence of $(u,v) \equiv (u^\delta, v^\delta)$ to its equilibrium state, $(\bar{u},\bar{v})$, governed by the limiting equations, $\bar{u}_t + \bar{v}_t + f(\bar{u})_x = 0$, $S(\bar{u},\bar{v}) = 0$. In particular, we provide sharp convergence rate estimates as the relaxation parameter $\delta \downarrow 0$. The novelty of our approach is the use of a weak $W^{-1}(L^1)$-measure of the error, which allows us to obtain sharp error estimates. It is shown that the error consists of an initial contribution of size $||S(u_0^\delta, v_0^\delta)||_{L^1}$, together with accumulated relaxation error of order $\mathcal{O}(\delta)$. The sharpness of our results is found to be in complete agreement with the numerical experiments reported in [Schroll, Tveito, and Winther, *SIAM J. Numer. Anal.*, 34 (1997), pp. 1152–1166].

Key words. conservation laws, stiff source terms, relaxation, *Lip*$^+$-stability, convergence rate estimates

AMS subject classifications. Primary, 35L65; Secondary, 58G16, 58G18

PII. S0036141096301488

1. Introduction.

We are concerned with one-dimensional systems of conservation laws which are coupled through a stiff source term. The purpose of this paper is to study a convergence rate of such systems to their equilibrium solutions as the stiff relaxation parameter tends to zero.

Our system takes the form

$$u_t + f(u)_x = -\frac{1}{\delta}S(u,v), \tag{1.1}$$

$$v_t = \frac{1}{\delta}S(u,v), \tag{1.2}$$

where $\delta > 0$ is the small relaxation parameter. The stiff source term, $S(u,v)$, and the convective flux, $f(u)$, are assumed to be smooth functions. We consider the Cauchy problem associated with (1.1)–(1.2), subject to periodic or compactly supported initial data

$$u(x,0) = u_0(x), \quad v(x,0) = v_0(x). \tag{1.3}$$

Here $u(x,t) := u^\delta(x,t)$, $v(x,t) := v^\delta(x,t)$ is the unique entropy solution of (1.1)–(1.3), which can be realized as the vanishing viscosity limit $u^\delta = \lim_{\nu\downarrow 0} u^{\delta,\nu}$, $v^\delta = \lim_{\nu\downarrow 0} v^{\delta,\nu}$, where $(u^{\delta,\nu}, v^{\delta,\nu})$ is the solution of the regularized viscosity system

$$u_t^{\delta,\nu} + f(u^{\delta,\nu})_x = -\frac{1}{\delta}S(u^{\delta,\nu}, v^{\delta,\nu}) + \nu u_{xx}^{\delta,\nu}\ , \tag{1.4}$$

$$v_t^{\delta,\nu} = \frac{1}{\delta}S(u^{\delta,\nu}, v^{\delta,\nu}). \tag{1.5}$$

* Received by the editors April 1, 1996; accepted for publication (in revised form) July 30, 1996. This research was supported by ONR grant N0014-91-J1076, NSF grant DMS91-03104, and GIF grant 318.195.06/93.

http://www.siam.org/journals/sima/28-6/30148.html

† School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel (kurganov@math.tau.ac.il).

‡ Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095 (tadmor@math.ucla.edu).

This regularized system, with fixed $\delta > 0$ (and $\nu > 0$), admits a unique, global (and, respectively, classical) solution. Indeed, such a solution can be constructed, for example, by fixed point iterations which alternate between the solution of the ODE (1.5) for v and the viscous conservation law—with v-dependent source term (1.4) for u. Moreover, by the maximum principle, e.g., [PW], the solution constructed admits a global uniform bound in view of our monotonicity assumption specified in section 2, $-S_u, S_v \leq 0$. Finally, by standard arguments (which we omit), there exists a constant, *independent* of ν, $C_\delta = \exp\{2(|S_u| + |S_v|)t/\delta\}$, such that

$$\begin{aligned}&\|u_1^\delta(\cdot,t) - u_2^\delta(\cdot,t)\|_{L^1} + \|v_1^\delta(\cdot,t) - v_2^\delta(\cdot,t)\|_{L^1}\\ &\quad\leq C_\delta\left[\|u_1^\delta(\cdot,0) - u_2^\delta(\cdot,0)\|_{L^1} + \|v_1^\delta(\cdot,0) - v_2^\delta(\cdot,0)\|_{L^1}\right].\end{aligned}$$

Consequently, the uniqueness of the viscous solution, $u^{\delta,\nu}$, and hence the uniqueness of its entropy limit BV-solution, u^δ, then follow. We refer to, e.g., [HW], [Lu], and [Le] for further discussions on the existence and uniqueness for various related models of the above type.

Once we identify the unique entropy solution, (u^δ, v^δ), we seek its equilibrium state as $\delta \downarrow 0$, $(\bar{u}, \bar{v})$. Formally, our equilibrium solution is governed by the limit system obtained by letting $\delta \downarrow 0$ in (1.1)–(1.2),

$$(\bar{u} + \bar{v})_t + f(\bar{u})_x = 0, \tag{1.6}$$

$$S(\bar{u}, \bar{v}) = 0. \tag{1.7}$$

To obtain the limiting equation (1.6), add (1.2) to (1.1); to obtain the constraint equation (1.7), multiply (1.2) by δ and pass to the formal limit as $\delta \to 0$.

The two main questions that we address in this paper are concerned with the convergence of the entropy solution (u^δ, v^δ) to its expected equilibrium state $(\bar{u}, \bar{v})$.

Convergence. We prove the convergence to the expected limits

$$\bar{u} = \lim_{\delta,\nu\downarrow 0} u^{\delta,\nu}, \qquad \bar{v} = \lim_{\delta,\nu\downarrow 0} v^{\delta,\nu}. \tag{1.8}$$

Moreover, we provide the following.

Error estimates. We estimate the convergence rate as $\nu \to 0$ and, in particular, as $\delta \to 0$.

Assume that $S_v \neq 0$ so we can solve the constraint equation (1.7) and obtain its solution in the explicit form

$$\bar{v} = v(\bar{u}). \tag{1.9}$$

Inserted into (1.6), we obtain that $\bar{u}$ is governed by the limiting equation

$$[\bar{u} + v(\bar{u})]_t + f(\bar{u})_x = 0. \tag{1.10}$$

Equivalently, if we denote $\bar{w} = \bar{w}(\bar{u}) := \bar{u} + v(\bar{u})$ and let its inverse[1] $\bar{u} = \bar{u}(\bar{w})$, then we conclude that the limiting equation (1.10) can be rewritten as a single conservation law, expressed in terms of the combined flux $F(\bar{w}) := f(\bar{u}(\bar{w}))$,

$$\bar{w}_t + F(\bar{w})_x = 0. \tag{1.11}$$

[1] The inverse exists since by our monotonicity assumption in section 2 below, $v'(u) = -S_u/S_v > -1$.

We obtain our convergence results under the assumptions of convexity—both $f(\cdot)$ and $F(\cdot)$ and the monotonicity of $S(u,v)$. In addition, we assume that we start with "prepared" initial data, in the sense that $u_0 \equiv u_0^\delta$ and $v_0 \equiv v_0^\delta$ approach their equilibrium state (1.7) as $\delta \downarrow 0$, i.e.,

$$||S(u_0^\delta(x), v_0^\delta(x))||_{L^1(x)} \xrightarrow{\delta\to 0} 0.$$

Specifically, we let $\epsilon = \epsilon(\delta) \downarrow 0$ denote the vanishing *initial error*

$$||S(u_0^\delta(x), v_0^\delta(x))||_{L^1(x)} \sim \epsilon(\delta) \downarrow 0. \tag{1.12}$$

Equipped with these assumptions, we formulate in section 2 our main results, which we summarize here in the following theorem.

THEOREM 1.1 (main theorem). *Consider the system* (1.3)–(1.5) *subject to* $W^2(L^1)$-*"prepared" initial data,* (1.12). *Then* $(u^{\delta,\nu}, v^{\delta,\nu})$ *converges to* $(\bar{u},\bar{v})$ *as* $\nu \to 0$, $\delta \to 0$, *and the following error estimate holds* $\forall p$, $1 \le p \le \infty$*:*

$$||u^{\delta,\nu}(\cdot,t) - \bar{u}(\cdot,t)||_{W^s(L^p(x))} \le \text{Const}_T \cdot \big(\epsilon(\delta) + \delta + \nu\big)^{\frac{1-sp}{2p}}, \quad -1 \le s \le \frac{1}{p}\ . \tag{1.13}$$

Thus, (1.13) reflects three sources for error accumulation: the initial error of size $\epsilon(\delta)$, the relaxation error of order δ, and the vanishing viscosity of order ν. For example, in the inviscid case ($\nu = 0$) and with "canonically prepared" initial data such that $\epsilon(\delta) \sim \delta$, we set $(s,p) = (0,1)$ in (1.13) to conclude an L^1-convergence rate of order $\mathcal{O}(\sqrt{\delta})$; in fact, in Corollary 2.3 below we extend this L^1-estimate to the v-variable, stating that

$$||u^\delta(\cdot,t) - \bar{u}(\cdot,t)||_{L^1} + \|v^\delta(\cdot,t) - \bar{v}(\cdot,t)\|_{L^1} = \mathcal{O}(\sqrt{\delta}). \tag{1.14}$$

The two-step proof of the main theorem is presented in sections 3 (stability) and 4 (consistency).

We close this introduction with three prototype examples.

Example 1: *Combustion.* We consider a combustion model proposed by Majda [Ma]. This model was consequently studied in [Le], [TY], and [Lu]. It takes the form

$$\begin{aligned} u_t + f(u)_x &= \frac{1}{\delta}A(u)v + \nu u_{xx}, \\ v_t &= -\frac{1}{\delta}A(u)v. \end{aligned} \tag{1.15}$$

Here $u \equiv u^{\delta,\nu}$ is a lumped variable representing some features of density, velocity, and temperature, while $v \equiv v^{\delta,\nu} \ge 0$ represents the mass fraction of unburnt gas in a simplified kinetics scheme; $\frac{1}{\delta}$ is the rate of reaction and the parameter $\nu > 0$ is a lumped parameter representing the effects of diffusion and heat conduction.

In this model, $S(u,v) = -A(u)v$ and our convexity and monotonicity assumptions (2.1)–(2.3) below hold, provided that

$$A'(u) < 0, \ A(u) \ge \eta > 0; \quad f''(u) \ge \alpha > 0. \tag{1.16}$$

The limiting equation (1.10) in this example reads

$$\bar{u}_t + f(\bar{u})_x = 0,$$

and hence $u^{\delta,\nu} - \bar{u}$ satisfies the error estimate (1.13).

Example 2: *Adsorption.* We consider the following stiff system:

$$\begin{aligned} u_t + f(u)_x &= -\frac{1}{\delta}(A(u) - v), \\ v_t &= \frac{1}{\delta}(A(u) - v). \end{aligned} \tag{1.17}$$

In this example $u \equiv u^\delta$ denotes the density of some species contained in a fluid flowing through a fixed bed, and $v \equiv v^\delta$ denotes the density of the species adsorbed on the material in the bed; $\delta > 0$ is referred to as the relaxation time. Different forms of adsorption functions, $A(u)$, are discussed in [STW], [TW1], [TW2], and the references therein.

The source term associated with this adsorption model, $S(u,v) = A(u) - v$, yields a limiting equation of the form

$$[\bar{u} + A(\bar{u})]_t + f(\bar{u})_x = 0.$$

Under the monotonicity assumption and convexity condition (consult (2.1)–(2.3)),

$$A'(u) \geq 0, \quad \left[\frac{f'(u)}{1 + A'(u)}\right]' \geq \alpha > 0. \tag{1.18}$$

We conclude the error estimate (1.13) with $\nu = 0$. In particular, for "canonically prepared" initial data such that $||A(u_0^\delta) - v_0^\delta||_{L^1} = \mathcal{O}(\delta)$, (1.14) yields a convergence rate of order $\mathcal{O}(\sqrt{\delta})$.

In this context it is interesting to contrast our above error estimates with those of [STW]. In [STW], Schroll, Tveito, and Winther studied the error estimates for the adsorption model (1.17) subject to "canonically prepared" initial data, $||A(u_0^\delta) - v_0^\delta||_{L^1} = \mathcal{O}(\delta)$, and concluded an L^1-convergence rate of order $\mathcal{O}(\delta^{\frac{1}{3}})$. Their reported numerical experiments, however, indicate a faster convergence rate of order $\mathcal{O}(\sqrt{\delta})$. Our results, e.g., (1.14), apply to their numerical experiments and confirm this optimal $\mathcal{O}(\sqrt{\delta})$ convergence rate. It should be pointed that the $\mathcal{O}(\delta^{\frac{1}{3}})$ error estimate in [STW] was derived by interpolation between L^2- and L^1-error bounds. It is here that we take advantage of our *sharper* interpolation between the *weaker* $\mathcal{O}(\delta)$ *Lip*$'$- and the $\mathcal{O}(1)$ BV-bounds. This enables us to improve over [STW] in both simplicity and generality and conclude with the sharper estimate of order $\mathcal{O}(\sqrt{\delta})$.

Example 3: *Relaxation.* Let us consider the following semilinear stiff system (see, e.g., [JX], [Li]):

$$\begin{aligned} u_t + v_x &= 0, \\ v_t + au_x &= \frac{1}{\delta}S(u,v), \end{aligned} \tag{1.19}$$

where $S(u,v) := f(u) - v$ and a is given positive number. The limiting equation, with $v(u) = f(u)$, is then

$$\bar{u}_t + f(\bar{u})_x = 0.$$

To study this system we rewrite it in the form of (1.1)–(1.2) by means of two changes of variables. First, we define the characteristic variables $w := \sqrt{a}\,u + v$, $z :=$

$\sqrt{a}\,u - v$. The system (1.19) then takes the form

$$\begin{aligned} z_t - \sqrt{a}\,z_x &= -\frac{1}{\delta}S(z,w), \\ w_t + \sqrt{a}\,w_x &= \frac{1}{\delta}S(z,w), \end{aligned} \tag{1.20}$$

with $S(z,w) = S(u(z,w), v(z,w))$. Next, we make the second change of variables, $x' := x - \sqrt{a}\,t$, obtaining

$$\begin{aligned} z_t - 2\sqrt{a}\,z_{x'} &= -\frac{1}{\delta}S(z,w), \\ w_t &= \frac{1}{\delta}S(z,w). \end{aligned} \tag{1.21}$$

In this model, the flux is linear and hence our first convexity assumption, (2.2), holds. The second one, (2.3), is satisfied for convex f's. In addition, the monotonicity of S, $S_z \geq 0, S_w \leq -\eta < 0$, amounts (in terms of S_u and S_v) to the inequalities

$$S_v \leq -\eta < 0, \quad S_v\sqrt{a} \leq S_u \leq -S_v\sqrt{a}.$$

Thus, $S(u,v) = f(u) - v$ should satisfy Liu's *subcharacteristic* condition (e.g., [Li]),

$$-\sqrt{a} \leq f'(u) \leq \sqrt{a}.$$

In this case, our main theorem with $p = 1$, for example, yields

$$||u^\delta - \bar{u}||_{W^s(L^1)} = \text{Const} \cdot \Big(||f(u_0^\delta) - v_0^\delta||_{L^1} + \delta\Big)^{\frac{1-s}{2}}, \quad -1 \leq s \leq 1.$$

2. Statement of main results. We seek the behavior of the solution of regularized system (1.4)–(1.5) towards the limit solution as $\delta \to 0$, as well as $\nu \to 0$. Throughout this section we make the following two main assumptions.

Monotonicity. $S(u,v)$ is monotonic with respect to u and strictly monotonic with respect to v,

$$S_u(u,v) \geq 0, \quad S_v(u,v) \leq -\eta < 0. \tag{2.1}$$

Convexity. $f(\cdot)$ is convex and $F(\cdot)$ is a strictly convex function,

$$f''(u) \geq 0, \tag{2.2}$$

$$F''(w) \geq \alpha > 0 \Longleftrightarrow \left(\frac{f'(\bar{u})}{1 + v'(\bar{u})}\right)' \geq \alpha > 0. \tag{2.3}$$

Remark. Our first assumption of monotonicity guarantees, by the classical maximum principle (see, e.g., [PW]), the L^∞-boundedness of $(u^{\delta,\nu}, v^{\delta,\nu})$ (proof is left to the reader).

Equipped with the two assumptions above, we now turn to the main result of this paper. To this end, our error estimate is formulated in terms of the *weak Lip'*-(semi)norm, $||\cdot||_{Lip'}$, and the dual of the *Lip*-norm given by

$$||\phi||_{Lip'} := \sup_{\psi}[(\phi - \hat{\phi_0}, \psi)/||\psi||_{W^{1,\infty}}], \quad \hat{\phi_0} := \int\limits_{supp\phi} \phi.$$

Thus, the *Lip*′-size of regular ϕ's (with bounded average over their finite support) amounts to their $W^{-1}(L^1)$-size or, equivalently, the L^1-size of their *primitive.* As we shall see, such weak (semi)norm has the advantage of providing us with *sharp* error estimates which, in turn, will be converted into strong ones.

THEOREM 2.1. *Consider the system* (1.3)–(1.5) *subject to* $W^2(L^1)$-*"prepared" initial data,* (1.12). *Then* $(u^{\delta,\epsilon}, v^{\delta,\epsilon})$ *converges to* $(\bar{u}, \bar{v})$ *as* $\delta \to 0$, $\nu \to 0$, *and the following error estimate holds:*

$$||u^{\delta,\nu}(\cdot,T) - \bar{u}(\cdot,T)||_{Lip'(x)} \leq \text{Const}_T \cdot \big(\epsilon(\delta) + \delta + \nu\big). \tag{2.4}$$

Let us consider the particular inviscid case, where $\nu = 0$. Then the entropy solution of the stiff system (1.1)–(1.2), (u^δ, v^δ), converges as $\delta \to 0$ to its equilibrium solution, $(\bar{u}, \bar{v})$, and we obtain the asserted convergence rate in terms of the initial error $\epsilon(\delta)$ and the vanishing relaxation parameter δ:

$$||u^{\delta}(\cdot,T) - \bar{u}(\cdot,T)||_{Lip'(x)} \leq \text{Const}_T \cdot \big(\epsilon(\delta) + \delta\big). \tag{2.5}$$

Remarks. 1. Our assumption of "prepared" initial data means that at the initial moment, $||S(u_0^\delta, v_0^\delta)||_{L^1} \overset{\delta\to 0}{\longrightarrow} 0$. In section 4 we will show that, in fact, $||S(u^{\delta,\nu}, v^{\delta,\nu})||_{L^1} \overset{\delta\to 0}{\longrightarrow} 0$ for all $t > 0$.

2. What about "nonprepared" initial data? In this case the initial layer formed persists in time; i.e., the initial error propagates and prevents convergence of $u^{\delta,\nu}, v^{\delta,\nu}$ to their equilibrium state.

The proof of the main theorem will be given in sections 3 and 4. To obtain this result we utilize the framework of Tadmor and Nessyahu [Ta], [NT]. To this end, we need the two ingredients of *consistency* and *stability.* Here, *consistency*—evaluated in terms of the *Lip*′-norm—measures by how much the approximate pair $(u^{\delta,\nu}, v(u^{\delta,\nu}))$ fails to satisfy the limiting equation (1.10); *stability* requires the *Lip*$^+$-stability [2] of $u^{\delta,\nu}$; that is, we seek a one-sided Lipschitz continuity (OSLC) of the viscosity solution $u^{\delta,\nu}$,

$$\|u^{\delta,\nu}(\cdot,t)\|_{Lip^+(x)} := \sup_x \,[u_x^{\delta,\nu}(x,t)]_+ \leq C_t \cdot \|u^{\delta,\nu}(\cdot,0)\|_{Lip^+(x)}. \tag{2.6}$$

By interpolation between the (weak) *Lip*′-error estimate (2.4) and the (strong) *BV*-boundedness of the error (which follows from the *Lip*$^+$-boundedness due to (2.6)), we are able to convert the weak error estimate stated in Theorem 2.1 into a *strong* one. As in [NT], we conclude with the following corollary.

COROLLARY 2.2 (global estimate). *Consider the inviscid problem* (1.1)–(1.3), (1.12). *Then the following convergence rate estimate holds:*

$$||u^{\delta}(x,T) - \bar{u}(x,T)||_{L^p} \leq \text{Const}_T \cdot (\epsilon(\delta) + \delta)^{\frac{1}{2p}}, \qquad 1 \leq p \leq \infty. \tag{2.7}$$

Remark. The above-mentioned L^p-estimates in (2.7) are, in fact, particular cases of the more general error estimate in the $W^s(L^p)$-norm

$$||u^{\delta}(x,T) - \bar{u}(x,T)||_{W^s(L^p)} \leq \text{Const}_T \cdot (\epsilon(\delta) + \delta)^{\frac{1-sp}{2p}}, \qquad -1 \leq s \leq \frac{1}{p}\ . \tag{2.8}$$

[2] Here $||\phi||_{Lip^+} := ess\sup_{x\neq y}\left[\frac{\phi(x)-\phi(y)}{x-y}\right]_+$, where, as usual, $(\cdot)_+$ denotes the "positive part of." For convenience we shall use the equivalent definition of the *Lip*$^+$ norm: $||\phi||_{Lip^+} := \sup_x\,[\phi'(x)]_+$, where the derivative of ϕ is taken in the distribution sense.

The special cases, $(s,p)=(-1,1)$ and $s=0$, correspond, respectively, to the weak *Lip*′-estimate (Theorem 2.1) and the global L^p-estimate (Corollary 2.2).

Taking $p=1$ in (2.7), we obtain, in particular, the L^1-error estimate, which reads

$$||u^\delta(x,T)-\bar{u}(x,T)||_{L^1} \leq \text{Const}_T \cdot \sqrt{\epsilon(\delta)+\delta}. \tag{2.9}$$

In this L^1-framework, we are able to extend the last estimate and obtain the same $\mathcal{O}(\sqrt{\epsilon(\delta)+\delta})$ convergence rate of v^δ towards $\bar{v}$. This brings us to the following corollary.

COROLLARY 2.3 (L^1-error estimate). *Consider the system* (1.1)–(1.3) *subject to "prepared" initial data,* (1.12)*. Then we have*

$$||u^\delta(x,T)-\bar{u}(x,T)||_{L^1}+||v^\delta(x,T)-\bar{v}(x,T)||_{L^1} \leq \text{Const}_T \cdot \sqrt{\epsilon(\delta)+\delta}. \tag{2.10}$$

In particular, for "canonically prepared" initial data, $||S(u_0^\delta,v_0^\delta)||_{L^1}=\epsilon(\delta)\sim\delta$, *we obtain a convergence rate of order* $\sqrt{\delta}$,

$$||u^\delta(x,T)-\bar{u}(x,T)||_{L^1}+||v^\delta(x,T)-\bar{v}(x,T)||_{L^1} \leq \text{Const}_T \cdot \sqrt{\delta}. \tag{2.11}$$

Proof. We first note that due to the strict monotonicity of $S(u,v)$ with respect to its second argument and the L^∞-bound of $u^\delta, v^\delta, \bar{u}$, and $\bar{v}$, we have

$$\begin{aligned} |v^\delta-\bar{v}| &= |v^\delta-v(u^\delta)+v(u^\delta)-\bar{v}| \leq |v^\delta-v(v^\delta)|+|v(v^\delta)-\bar{v}| \\ &= |v'(\tilde{u})|\cdot|u^\delta-\bar{u}|+\left|\frac{S(u^\delta,v^\delta)-S(u^\delta,v(u^\delta))}{S_v(u^\delta,\tilde{v})}\right| \sim |u^\delta-\bar{u}|+|S(u^\delta,v^\delta)|. \end{aligned}$$

Here $\tilde{u}$ and $\tilde{v}$ are appropriate midvalues, $\tilde{u}=\theta_1 u^\delta+(1-\theta_1)\bar{u}$, $\tilde{v}=\theta_2 v^\delta+(1-\theta_2)\bar{v}$. And we now obtain the desired estimate, (2.10),

$$\begin{aligned} ||v^\delta(x,T)-\bar{v}(x,T)||_{L^1} &\leq \text{Const}_T\cdot(||u^\delta(x,T)-\bar{u}(x,T)||_{L^1} \\ &\quad +||S(u^\delta(x,T),v^\delta(x,T))||_{L^1}) \\ &= \mathcal{O}(\sqrt{\epsilon(\delta)+\delta})+\mathcal{O}(\epsilon(\delta)+\delta)=\mathcal{O}(\sqrt{\epsilon(\delta)+\delta}). \end{aligned} \tag{2.12}$$

Indeed, the first $\mathcal{O}(\sqrt{\bullet})$-upperbound on the right is due to (2.9); the second upperbound, $||S(u^\delta(x,T),v^\delta(x,T))||_{L^1}=\mathcal{O}(\epsilon(\delta)+\delta)$, is outlined in section 4 below. □

Finally, arguing along the lines of [NT; Corollary 2.4], we also obtain the *pointwise* convergence towards the equilibrium solution away from discontinuities.

COROLLARY 2.4 (local estimate). *Consider the inviscid problem* (1.1)–(1.3), (1.12)*. Then the following estimate holds:*

$$|u^\delta(x,T)-\bar{u}(x,T)| \leq \text{Const}_{x,T}\cdot(\epsilon(\delta)+\delta)^{\frac{1}{3}} . \tag{2.13}$$

Here, $\text{Const}_{x,T}$ *is a constant which measures the local smoothness of* $u(\cdot,T)$ *in the small neighborhood of* x,

$$\text{Const}_{x,T} \sim 1+\max_{|y-x|<\sqrt[3]{\delta}}|\bar{u}_x(y,T)|.$$

.

3. *Lip*$^+$-stability estimate. We now turn to the proof of our main theorem. We begin with the *Lip*$^+$-stability of the solution of (1.4)–(1.5).

Assertion 3.1. *Consider the system* (1.4), (1.5) *subject to Lip*$^+$*-bounded initial data* (1.3). *Then there exists a constant (which may depend on the initial data) such that*

$$||u^{\delta,\nu}(\cdot,T)||_{Lip^+(x)} \leq \text{Const.} \tag{3.1}$$

Proof. The proof is based on the maximum principle for $(u_x^{\delta,\nu})_+$.

Differentiation of (1.4) and (1.5) with respect to x implies

$$(u_x^{\delta,\nu})_t + f''(u^{\delta,\nu})(u_x^{\delta,\nu})^2 + f'(u^{\delta,\nu})(u_x^{\delta,\nu})_x = -\frac{1}{\delta}[S_u u_x^{\delta,\nu} + S_v v_x^{\delta,\nu}] + \nu(u_x^{\delta,\nu})_{xx}, \tag{3.2}$$

$$(v_x^{\delta,\nu})_t = \frac{1}{\delta}[S_u u_x^{\delta,\nu} + S_v v_x^{\delta,\nu}]. \tag{3.3}$$

We now multiply (3.2) by $\frac{1+sgn(u_x^{\delta,\nu})}{2}$; using the monotonicity of $S(u,v)$ and convexity of $f(u)$ we obtain the following inequalities:

$$\begin{aligned}&[(u_x^{\delta,\nu})_+]_t + f'(u^{\delta,\nu}) \cdot [(u_x^{\delta,\nu})_+]_x \\ &\qquad \leq -\frac{1}{\delta}\left[S_u(u_x^{\delta,\nu})_+ + S_v v_x^{\delta,\nu}\left(\frac{1+sgn(u_x^{\delta,\nu})}{2}\right)\right] + \nu[(u_x^{\delta,\nu})_+]_{xx},\end{aligned} \tag{3.4}$$

$$(v_x^{\delta,\nu})_t \leq \frac{1}{\delta}[S_u(u_x^{\delta,\nu})_+ + S_v v_x^{\delta,\nu}]. \tag{3.5}$$

By solving the second inequality, we find (with $S_v(\tau) := S_v(x,\tau) \equiv S_v(u^{\delta,\nu}(x,\tau), v^{\delta,\nu}(x,\tau))$ and $B(t) := \int_0^t S_v(\tau)d\tau$) that

$$v_x^{\delta,\nu}(t) \leq e^{\frac{B(t)}{\delta}} v_x^{\delta,\nu}(0) + \frac{1}{\delta}\int_0^t e^{\frac{B(t)-B(\tau)}{\delta}} S_u(\tau)(u_x^{\delta,\nu}(\tau))_+ d\tau. \tag{3.6}$$

Plugging this into (3.4) and denoting $m(t) = \max_x (u_x^{\delta,\nu}(x,t))_+$, we end up with

$$\dot{m}(t) \leq -\frac{S_u(t)}{\delta}m(t) - \frac{S_v(t)}{\delta}e^{\frac{B(t)}{\delta}}(v_x^{\delta,\nu}(0))_+ - \frac{S_v(t)}{\delta^2}\int_0^t e^{\frac{B(t)-B(\tau)}{\delta}} S_u(\tau)m(\tau)d\tau. \tag{3.7}$$

The first and the third terms in the right-hand side of (3.7) add up to a perfect derivative, modulo extra terms which are differentiated along the characteristics where $u_{xx}^{\delta,v}(x(t),t) = 0$ so that

$$\dot{m}(t) \leq \left(-e^{\frac{B(t)}{\delta}}(v_x^{\delta,\nu}(0))_+\right)_t - \frac{1}{\delta}\left(\int_0^t e^{\frac{B(t)-B(\tau)}{\delta}} S_u(\tau)m(\tau)d\tau\right)_t + k(t). \tag{3.8}$$

Here the constant $k(t)$ (depending on the convexity constant of F in (2.3), α) is an upperbound on the extra terms differentiated along the characteristics, e.g., $\partial_x B(x,t)\dot{x}e^{B(t)/\delta}(u_x^{\delta,v}(0))_+/\delta \dots$. Integration of (3.8) over $(0,T)$ yields

$$m(T) \leq m(0) + (v_x^{\delta,\nu}(0))_+\left[1 - e^{\frac{B(T)}{\delta}}\right] - \frac{1}{\delta}\int_0^T e^{\frac{B(T)-B(\tau)}{\delta}} S_u(\tau)m(\tau)d\tau + \int_0^T k(\tau)d\tau. \tag{3.9}$$

In view of the positivity of S_u, we obtain that

$$(u_x^{\delta,\nu}(x,T))_+ \leq (u_x^{\delta,\nu}(x,0))_+ + (v_x^{\delta,\nu}(x,0))_+ + K_T, \quad K_T = \int_0^T k(\tau)d\tau,$$

and the assertion follows with $\text{Const} = ||u^{\delta,\nu}(\cdot,0)||_{Lip^+(x)} + ||v^{\delta,\nu}(\cdot,0)||_{Lip^+(x)} + K_T$. □

We close this section by noting that the proof of Assertion 3.1 is based on the straightforward, formal maximum principle for the positive part of $u^{\delta,\nu}$; alternatively, it could be justified, for example, by L^p iterations in (3.4).

4. *Lip*'-consistency and proof of the main result. In this section we prove the promised error estimate (2.4) in the Lip'-norm. According to the results of [Ta], [NT], the error $||u^{\delta,\nu} - \bar{u}||_{Lip'}$ is upper bounded by the truncation error

$$\left|\left|[u^{\delta,\nu} + v(u^{\delta,\nu})]_t + f(u^{\delta,\nu})_x\right|\right|_{Lip'(x,t)}. \tag{4.1}$$

This quantity measures by how much $u^{\delta,\nu}$ fails to satisfy the limiting equation (1.10). To complete this proof we have to show, therefore, that the truncation error is of order $\mathcal{O}(\epsilon(\delta) + \delta + \nu)$. We proceed as follows.

Adding the two components of the regularized system (1.5) to (1.4), we obtain that

$$u_t^{\delta,\nu} + v_t^{\delta,\nu} + f(u^{\delta,\nu})_x = \nu u_{xx}^{\delta,\nu},$$

which we rewrite as

$$\begin{aligned}*[u^{\delta,\nu} + v(u^{\delta,\nu})]_t + f(u^{\delta,\nu})_x &= u_t^{\delta,\nu} + v_t^{\delta,\nu} + f(u^{\delta,\nu})_x \\ &+[v(u^{\delta,\nu}) - v^{\delta,\nu}]_t = \nu u_{xx}^{\delta,\nu} + [v(u^{\delta,\nu}) - v^{\delta,\nu}]_t.\end{aligned}$$

It is here that we take advantage of the *weak* Lip'-norm introduced earlier in section 2: by measuring the L^1-size of its primitive, the right-hand side of the last equality tells us that the truncation error in (4.1) does not exceed

$$\begin{aligned}&||\nu u_{xx}^{\delta,\nu} + [v(u^{\delta,\nu}) - v^{\delta,\nu}]_t||_{Lip'(x,t)} \\ &\qquad \leq \text{Const}_T \cdot \left[\nu ||u_x^{\delta,\nu}||_{L^1(x,t)} + ||v(u^{\delta,\nu}) - v^{\delta,\nu}||_{L^1(x,t)}\right] \\ &\qquad =: \text{Const}_T \cdot \left[\ I + II\ \right].\end{aligned} \tag{4.2}$$

We proceed with estimating the two terms on the right. First, since $u^{\delta,\nu}$ is Lip^+-bounded, (3.1), it has a bounded variation, $||u_x^{\delta,\nu}||_{L^1(x,t)} \leq C_K$ (where C_K may depend on the Lip^+-bound, K, and the finite support of $u^{\delta,\nu}$) and, therefore, $I \leq \mathcal{O}(\nu)$. Next, we find that the second term, II, is of order

$$II \equiv ||v(u^{\delta,\nu}) - v^{\delta,\nu}||_{L^1(x,t)} \sim ||S(u^{\delta,\nu}, v^{\delta,\nu})||_{L^1(x,t)}. \tag{4.3}$$

Indeed, since $0 < \eta \leq -S_v \leq \text{Const}$, we have

$$\frac{1}{\eta} \leq \frac{|v(u^{\delta,\nu}) - v^{\delta,\nu}|}{|S(u^{\delta,\nu}, v(u^{\delta,\nu})) - S(u^{\delta,\nu}, v^{\delta,\nu})|} \leq \text{Const}$$

and, hence, $|v(u^{\delta,\nu}) - v^{\delta,\nu}| \sim |S(u^{\delta,\nu}, v(u^{\delta,\nu})) - S(u^{\delta,\nu}, v^{\delta,\nu})| = |S(u^{\delta,\nu}, v^{\delta,\nu})|$, and (4.3) follows. Returning to (4.2) we find that

$$\begin{aligned} ||\nu u^{\delta,\nu}_{xx} + [v(u^{\delta,\nu}) - v^{\delta,\nu}]_t||_{Lip'(x,t)} &\leq \text{Const}_T \cdot \Big[\ I + II\ \Big] \\ &\leq \text{Const}_T \cdot \Big[\nu + ||S(u^{\delta,\nu}, v^{\delta,\nu})||_{L^1(x,t)}\Big]. \end{aligned} \tag{4.4}$$

To conclude with the promised $\mathcal{O}(\epsilon(\delta) + \delta + \nu)$-bound, it remains to prove that $||S(u^{\delta,\nu}, v^{\delta,\nu})||_{L^1(x,t)}$—or, utilizing (4.3), that $\delta||v^{\delta,\nu}_t(\cdot,t)||_{L^1(x)}$ is of order $\mathcal{O}(\epsilon(\delta)+\delta)$,

$$||S(u^{\delta,\nu}(\cdot,t), v^{\delta,\nu}(\cdot,t))||_{L^1(x)} \equiv \delta||v^{\delta,\nu}_t(\cdot,t)||_{L^1(x)} = \mathcal{O}(\epsilon(\delta) + \delta). \tag{4.5}$$

To achieve such an estimate, we differentiate (1.4) with respect to t, multiply by $sgn(u^{\delta,\nu}_t)$, and obtain

$$\begin{aligned} |u^{\delta,\nu}_t|_t + (f'(u^{\delta,\nu})u^{\delta,\nu}_t)_x sgn(u^{\delta,\nu}_t) = -\frac{1}{\delta}\Big(S_u|u^{\delta,\nu}_t| + S_v|v^{\delta,\nu}_t|sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t)\Big) \\ +\epsilon(u^{\delta,\nu}_t)_{xx} sgn(u^{\delta,\nu}_t). \end{aligned} \tag{4.6}$$

The same treatment of equation (1.5) yields

$$|v^{\delta,\nu}_t|_t = \frac{1}{\delta}\Big(S_u|u^{\delta,\nu}_t|sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t) + S_v|v^{\delta,\nu}_t|\Big). \tag{4.7}$$

Next, we integrate the following equations with respect to x:

$$\frac{d}{dt}||u^{\delta,\nu}_t||_{L^1(x)} \leq -\frac{1}{\delta}\left(\int\limits_x S_u|u^{\delta,\nu}_t|dx + \int\limits_x S_v|v^{\delta,\nu}_t|sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t)dx\right), \tag{4.8}$$

$$\frac{d}{dt}||v^{\delta,\nu}_t||_{L^1(x)} \leq \frac{1}{\delta}\left(\int\limits_x S_u|u^{\delta,\nu}_t|sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t)dx + \int\limits_x S_v|v^{\delta,\nu}_t|dx\right). \tag{4.9}$$

Finally, we add up (4.8) and (4.9), obtaining

$$\begin{aligned} \frac{d}{dt}\Big[||u^{\delta,\nu}_t||_{L^1(x)} + ||v^{\delta,\nu}_t||_{L^1(x)}\Big] \leq \frac{1}{\delta}\Bigg[&\int\limits_x S_u|u^{\delta,\nu}_t|\Big(sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t) - 1\Big)dx \\ &+ \int\limits_x S_v|v^{\delta,\nu}_t|\Big(1 - sgn(u^{\delta,\nu}_t)sgn(v^{\delta,\nu}_t)\Big)dx\Bigg] \leq 0. \end{aligned}$$

It follows that

$$||u^{\delta,\nu}_t(\cdot,t)||_{L^1(x)} + ||v^{\delta,\nu}_t(\cdot,t)||_{L^1(x)} \leq ||u^{\delta,\nu}_t(\cdot,0)||_{L^1(x)} + ||v^{\delta,\nu}_t(\cdot,0)||_{L^1(x)} \tag{4.10}$$

and, in particular,

$$\delta||v^{\delta,\nu}_t(\cdot,t)||_{L^1(x)} \leq \delta||u^{\delta,\nu}_t(\cdot,0)||_{L^1(x)} + \delta||v^{\delta,\nu}_t(\cdot,0)||_{L^1(x)}.$$

To conclude this proof, we show that the upper bound on the right does not exceed the promised $\mathcal{O}(\epsilon(\delta)+\delta)$. Indeed, by equations (1.4)–(1.5), $u^{\delta,\nu}_t = -v^{\delta,\nu}_t - f(u^{\delta,\nu})_x + \nu u^{\delta,\nu}_{xx}$, and hence

$$\begin{aligned} \delta||u^{\delta,\nu}_t(\cdot,0)||_{L^1(x)} + \delta||v^{\delta,\nu}_t(\cdot,0)||_{L^1(x)} \leq{}& 2||S(u^{\delta,\nu}(\cdot,0), v^{\delta,\nu}(\cdot,0))||_{L^1(x)} \\ &+\delta||f(u^{\delta,\nu}(\cdot,0))_x||_{L^1(x)} + \delta\nu||u^{\delta,\nu}_{xx}(\cdot,0)||_{L^1(x)}. \end{aligned}$$

The three terms on the right are upper-bounded by $\mathcal{O}(\epsilon(\delta)+\delta)$ since, by our assumption of the "prepared" initial data (1.12), $||S(u^{\delta,\nu}(\cdot,0),v^{\delta,\nu}(\cdot,0))||_{L^1(x)}=\mathcal{O}(\epsilon(\delta))$; the BV-boundedness of $u^{\delta,\nu}$ yields $\delta||f(u^{\delta,\nu}(\cdot,0))_x||_{L^1(x)}=\mathcal{O}(\delta)$ and, finally, since the initial data are assumed to be in $W^2(L^1)$, then $\delta\nu||u^{\delta,\nu}_{xx}(\cdot,0)||_{L^1(x)}=\mathcal{O}(\delta\nu)<<\mathcal{O}(\delta)$. This completes the proof of Theorem 2.1. □

Remark. We close by noting that the $W^2(L^1)$-regularity of initial data used in the last stage of the proof can be relaxed. In fact, it is sufficient to assume $||u_{0x}||_{L^1}+\nu||u_{0xx}||_{L^1}\leq$ Const.

Note added in proof. We thank Professor R. Natalini for pointing out a gap in the previous version of the proof of Assertion 3.1. Details will appear elsewhere.

REFERENCES

[HW] J. Hu and J. Wang, *The Cauchy problem for a special system of quasilinear equations*, J. Partial Differential Equations, 4 (1991), pp. 33–44.

[JX] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[Le] A. Levy, *On Majda's model for dynamic combustion*, Comm. Partial Differential Equations, 17 (1992), pp. 657–698.

[Li] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[Lu] Y. Lu, *Cauchy problem for an extended model of combustion*, Proc. Royal Soc. Edinburgh Sect. A, 120 (1992), pp. 349–360.

[Ma] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70–93.

[NT] 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.

[PW] M. H. Protter and H. F. Weinberger, *Maximum Principles in DEs*, Springer-Verlag, Berlin, New York, 1967.

[STW] 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. 136–161.

[Ta] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 811–906.

[TW1] 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.

[TW2] A. Tveito and R. Winther, *An error estimate for a finite difference scheme approximating a hyperbolic system of conservation laws*, SIAM J. Numer. Anal., 30 (1993), pp. 401–424.

[TY] Z. Teng and L. Ying, *Existence, uniqueness and convergence as vanishing viscosity for a reaction-diffusion-convection system*, Acta Math. Sinica (N.S.), 5 (1989), pp. 114–135.

SIAM J. MATH. ANAL.
Vol. 28, No. 6, pp. 1457–1489, November 1997

DYNAMICS AND CONDENSATION OF COMPLEX SINGULARITIES FOR BURGERS' EQUATION I[*]

DAVID SENOUF[†]

Dedicated to Sophia de Rosnay

Abstract. Spatial analyticity properties of the solution to Burgers' equation with a generic initial data are presented, following the work of Bessis and Fournier [*Research Reports in Physics: Nonlinear Physics*, Springer-Verlag, Berlin, Heidelberg, 1990, pp. 252–257]. The positive viscosity solution is a meromorphic function with a countable set of conjugate poles confined to the imaginary axis. Their motion is governed by an infinite-dimensional Calogero dynamical system (CDS). The inviscid solution is a three-sheeted Riemann surface with three branch-point singularities.

Exact pole locations are found independent of the viscosity at the inviscid shock time t_*. For $t \neq t_*$, the time evolution of the poles is obtained numerically by solving a truncated version of the CDS. A Runge–Kutta scheme is used together with a "multipole" algorithm to deal with the computationally intensive nonlinear interaction of the poles. Additionally, for $t \leq t_*$, the small viscosity behavior of the poles is shown to be a perturbation of the conjugate inviscid branch-point singularities $\pm x_s(t)$. The numerical pole dynamics also provide the width of the analyticity strip which remains uniformly bounded away from zero, agreeing with asymptotic predictions.

For small $\nu > 0$ and $t \geq t_*$, different saddle-point approximations of the solution are found within and outside the caustics $x = \pm x_s(t)$. The transition between the two regimes at $x = \pm x_s(t)$ is described by a uniform asymptotic expansion involving the Pearcey integral. The solution is computed for small viscosity using pole dynamics, finite differences (method of lines), and asymptotic methods (saddle-point method); numerical agreement is established.

Key words. partial differential equations, asymptotic approximations, pole dynamics, domain of analyticity

AMS subject classifications. 35A20, 35A40, 35B40, 35Q53, 41A60

PII. S0036141095289373

1. Introduction.

In this article we investigate the spatial analyticity properties of a solution to Burgers' equation (hereafter referred to as "BE"):

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \nu\,\frac{\partial^2 u}{\partial x^2}, \qquad x \in \mathbb{R},\ t > 0,\ \nu \geq 0, \tag{1.1}$$

where the parameter ν is a viscosity coefficient and $u = u_\nu(x,t)$ represents the velocity field of a fluid particle at position x in space and time t. BE is a model for the statistical theory of turbulence [9, 10] which can be thought of as a simplified one-dimensional scalar analog of the Navier–Stokes equations of fluid dynamics. Although it does not exhibit the complexity of the Navier–Stokes equations, it does illustrate the interaction between a nonlinear first-order convective term and a second-order diffusive one. This feature, which is shared with the Navier–Stokes equations, may help us to understand the questions of regularity of so complicated a system. Although Burgers' model is not a good model for turbulence because it does not exhibit any chaotic behavior, there are other important applications besides the Ising model

[*] Received by the editors July 24, 1995; accepted for publication (in revised form) June 25, 1996. This research was partially supported by NSF grant DMS-9306720.

http://www.siam.org/journals/sima/28-6/28937.html

[†] Lehman Brothers, One Broadgate, London EC2M 7HA, United Kingdom (dsenouf@lehman.com).

analogy presented by Bessis and Fournier in [7]. A partial list of such applications can be found in the introduction of the work by She, Aurell, and Frisch in [30].

There are many tools which can be used to evaluate domains of analyticity of solutions to nonlinear PDEs: Painlevé expansions, Padé approximants, pole dynamics, spectral methods, the abstract Cauchy–Kowalewski theorem, and, more generally, methods involving analytic norms.

The work of Bardos and Benachour [5] was pioneering because it used complex analytic techniques (analytic norms) to describe domains of regularity for equations of fluid dynamics: Bardos and Benachour showed that the loss of analyticity for the incompressible Euler equations in $\mathbb{R}^n$ follows from a blowup in the vorticity $\omega = \nabla \times u$, in analogy with the blowup of the solution of the inviscid BE ($\nu = 0$), which is driven by the blowup of the gradient of the solution $\partial u/\partial x$.

Another method which also involves analytic norms is the abstract Cauchy–Kowalewski theorem. A concise and improved version of the original work of Nirenberg [27] can be found in [11]. One of the limitations of this method is that it cannot deal with parabolic PDEs. An interesting combination of Painlevé expansions and the abstract Cauchy–Kowalewski theorem applied to BE can be found in [23].

As far as spectral methods are concerned, we only mention the works of Sulem, Sulem, and Frisch [35] and Fournier and Frisch [19], both of which deal with BE. Reference [35] focuses on the numerical implementation of spectral methods to evaluate the widths of analyticity strips of solutions to nonlinear PDEs. Reference [19] focuses on spectral methods applied to the deterministic and statistical BE. More references can be found in both of these works.

The method of pole dynamics originated with Kruskal's work [26], followed by Calogero [13] and the Choodnovskys [16], who developed a mathematical tool for what was to become a very powerful method to solve nonlinear PDEs. Indeed, they showed that a very large class of nonlinear evolution equations has an associated/equivalent N-body problem/formulation. This method consists in inserting into the PDE a pole expansion of the solution (a Mittag–Leffler expansion where the complex spatial poles are time dependent). Compatibility conditions are found in the form of a dynamical system for these poles; this dynamical system is referred to as the Calogero dynamical system (CDS for short).

In [20], Frisch and Morf describe complex time singularities for nonlinear PDEs as well as make a first attempt to describe spatial singularities for BE through pole dynamics. In [20, section 4], a list of other references which use pole dynamics can be found. In [19], Fournier and Frisch devoted greater attention to (real time) complex spatial singularities for the inviscid BE from a deterministic and a statistical point of view. In the meantime, Thual, Frisch, and Hénon [36] used pole dynamics to compute the stationary pole distribution and stationary solution of a Sivashinsky-type flame-front propagation pseudo-differential equation. From the work of Fournier and Frisch [19], Bessis and Fournier [6, 7] went on to analyze the spatial analytic properties for both the inviscid and viscous deterministic case using a generic initial data. Kimura [25] has described complex space and time pole positions for the BE with periodic initial data by straightforwardly solving for the roots of the Cole–Hopf variable.

In [6, 7], Bessis and Fournier studied the analytical properties of the solution to the BE with a cubic initial data, namely,

$$u(x,0) = u_0(x) = 4x^3 - x/t_*, \qquad x \in \mathbb{R}, \tag{1.2}$$

and t_* is a fixed positive parameter corresponding to the first blowup time of the solution to the inviscid equation. This work stemmed from the observation [19] that the large wavenumber asymptotic expansion of the Fourier transform of the inviscid solution was degenerate at the shock time t_*; it was even incorrect beyond t_*. Thus, Bessis and Fournier sought for an explanation using complex analytic methods. They showed that the inviscid shock could be interpreted as a permutation of the two sides of the "physical Riemann sheet." As far as the viscous case is concerned, Bessis and Fournier showed that the poles were confined to the imaginary axis and that they satisfied a CDS. Additionally, they presented a "limiting" pole density which characterized the process of pole condensation as $\nu \to 0^+$.

The inviscid equation $u_t + u\,u_x = 0$ is a simple hyperbolic quasi-linear PDE. Its solution develops a cube root singularity at the origin at the time $t = -(\inf_x u_0'(x))^{-1} > 0$, which, for (1.2), equals t_*. This is known to be a generic singularity for the inviscid BE. It is due to the coalescence at the origin ($x = 0$) of two complex conjugate branch points $\pm x_s(t)$ of order two. Thus, the cubic initial data is considered generic due to the local cube root shape of the shock of the inviscid solution at $t = t_*$ and the associated cube root singularity (for further details, see [32, App. A] and [7, 12, 19]). Another compelling reason for which a cubic polynomial is used is that its solution can be completely analyzed for both $\nu > 0$ and $\nu = 0$, unlike a higher-order polynomial of the form $u_0(x) = 2nx^{2n-1} - x/t_*$ (see [32, App. D]).

For $t > t_*$, the inviscid solution has three real values within the real interval $(-x_s, x_s)$, and in the real complement $(-x_s, x_s)^c$ it has one real value and two complex values (see [32, App. C] and [12]). By extending (for all $t > 0$) the domain of the spatial variable x and the range of the solution u into the complex plane, Bessis and Fournier have shown in [6, 7] that the analytic structure (topology) of the inviscid solution is a three-sheeted Riemann surface with three branch points. One is at infinity, and the other two come down the imaginary axis as a conjugate pair and coalesce at the origin at the shock time t_* to form a third-order branch point. The inviscid shock in the real plane is interpreted as the permutation of the physical Riemann sheets which make up the Riemann surface. More precisely, it appears to be the connecting path between the two sides of the physical Riemann sheet which are separated by nonphysical ones (see [6] for more details).

In this work, we propose to correct and extend the results of Bessis and Fournier in [7] by using complex analytic methods and asymptotic methods (including spectral methods). Most important, we show how to use pole dynamics to determine the evolution of the domain of regularity of the solution to BE. This method can be adapted to a wide range of nonlinear PDEs.

The positive viscosity solution ($\nu > 0$) is a meromorphic function with a countable set of conjugate pairs of simple poles for all $t > 0$. These poles move towards the origin along the imaginary axis, then turn around after a finite time and start moving away from the origin (see [32, Figs. 1.1, 1.2]). In the same way that the dynamics of the branch points of the inviscid solution help in understanding the formation of a shock in the real (physical) plane, we intend to illustrate the preservation of regularity of the viscous solution by further analyzing the dynamics of the simple poles. In turn, this will shed some light on the interaction between the nonlinear convective term and the diffusive one present in the (viscous) BE. Furthermore, we clarify the limiting process which describes the vanishing viscosity limit by focusing on the small ν asymptotic behavior of the poles. As $\nu \to 0^+$, the poles condense on the imaginary axis, yielding an asymptotic pole density. The inviscid limit can be recovered by introducing an integral representation of the Mittag–Leffler expansion which involves this density.

These results are presented in [32].

We focus on BE with the same cubic initial data (1.2) used by Bessis and Fournier, whose choice will be clarified. It should be noted, however, that the method of pole dynamics used in this article is adaptable to a large class of one-dimensional evolution equations (cf. [13, 16, 36]). The main difficulty in adapting this method to equations with spatial derivatives of order higher than two is that it may translate into additional algebraic conditions. Thus, one may have to solve a differential-algebraic system of equations (DAE). Additionally, the choice of the initial data need not be fixed and is actually the focus of current research in the case of BE with random initial data. Much attention has recently been devoted to this problem (see, for example, [4, 19, 26, 30, 34]). A final note on the generality of the method: the pole expansion and pole dynamics which are derived for BE are valid for any meromorphic solution to BE and as such correct the pole expansion previously derived for BE.

In part I, we describe exact and asymptotic properties of the positive viscosity solution, its pole locations, and their dynamics.

In section 2, the solution is explicitly given by the Cole–Hopf transform for $\nu > 0$. From a careful analysis of the Cole–Hopf variable, the solution is expressed in terms of its polar singularities by means of a Mittag–Leffler (pole) expansion. A correction to the infinite dimensional CDS derived by Bessis and Fournier which governs the time evolution of the poles is found by replacing the pole expansion of the solution into the PDE. This system represents compatibility conditions for the existence of such a pole expansion.

In section 3, from the integral representation obtained via the Cole–Hopf analysis and by means of the saddle-point method, we derive an asymptotic formula for the solution $u_\nu(x,t)$ for small ν. However, for $t \geq t_*$, there is a degeneracy in the asymptotic formula at the caustic $x = \pm x_s(t)$ where two saddle points coalesce; thus, we derive both the regular saddle-point analysis within and outside the caustic and a uniformly valid expansion via Pearcey's integral, which correctly describes the transition between the two regions. The asymptotic behavior of the solution at the caustic $u_\nu(x_s(t),t)$ is obtained from the Pearcey representation and is shown to match the behavior obtained from the classical saddle-point analysis.

In section 4, we analyze the pole locations: at the inviscid shock time t_* we use the Cole–Hopf variable to approximate the poles. Highly accurate asymptotic formulas of a related Fourier integral derived in [31] enable us to obtain almost exact pole locations independent of the viscosity. At other times ($t \neq t_*$), when no such formula can be obtained, we derive weaker asymptotic results: for small $\nu > 0$ and $0 < t \leq t_*$, we show that the poles are a perturbation of the inviscid branch-point singularities of the form $\beta_k(t,\nu) = |x_s(t)| + \mathcal{O}((k\nu)^{3/4})$. For $t > t_*$, their asymptotic behavior no longer depends on the inviscid branch-point singularities, and it is given by $\beta_k(t,\nu) = \mathcal{O}((k\nu)^{3/4})$. Similarly, we also show that for large k, fixed ν, and all $t > 0$, we have $\beta_k(t,\nu) = \mathcal{O}((k\nu)^{3/4})$.

In section 5, we analyze the time evolution of the poles more explicitly since their actual location for $t \neq t_*$ has not been described. The method consists in numerically solving a truncated version of the CDS. The "initial data" which is adjoined to this truncated system is generated by the exact pole locations found in section 4. A Runge–Kutta–Fehlberg 4–5 time marching scheme is used in combination with the "multipole" algorithm designed by Greengard and Rokhlin [21]. This multipole algorithm reduces the computational complexity of the nonlinear interaction of the poles in the Calogero ODE system from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$ particles (poles), thereby allowing us to carry out very large simulations with up to $N = 50,000$ poles. The closed-form solution to a two-pair pole dynamics is obtained and serves as a test case for the multipole simulation. The poles $\pm\beta_k(t,\nu)$ are confined to the imaginary axis and move towards the origin until a time $t = t_u(k)$, $k \in \mathbb{N}^*$; this is the time at which they turn around and move away from the origin. These turn-around times $t_u(k)$ decrease as k increases: $t_u(1) > t_u(2) > \cdots > t_u(n) > \cdots > 0$. Moreover, $t_u(1)$ occurs before t_* for $\nu \gtrapprox .01$ and after t_* for $\nu \lessapprox .01$. Another worthwhile feature is that t_u increases with decreasing ν, in accordance with the fact that the time at which the solution starts decaying increases with decreasing ν.

From this procedure, the evolution of the width of the analyticity strip is shown to remain uniformly bounded away from zero, agreeing with the asymptotic predictions and the well-known fact that BE with analytic initial data has a smooth solution for all times as long as $\nu > 0$ (in agreement with the results of Sulem, Sulem, and Frisch in [35]).

Finally, the solution is computed for small viscosity using pole dynamics, finite differences, and asymptotic methods (saddle-point analysis), and numerical agreement is established. The difference scheme we use is the method of lines consisting of the same Runge–Kutta–Fehlberg 4–5 scheme in time combined with central differencing in space. The solution is reconstructed from the pole positions and the Mittag–Leffler (pole) expansion of the solution.

In part II [32], the zero-viscosity limit of the solution is obtained via a process of pole condensation. It is shown that the asymptotic density of poles, which describes their condensation on the imaginary axis, can be obtained as the weak limit of a discrete Borel measure (analogously to the zero-dispersion limit of the spectral measure in the KdV problem). The analytic structure of the inviscid solution, which is a three-sheeted Riemann surface with three branch-point singularities, is recovered. The continuum limit of the pole expansion of the solution and the CDS for the poles is a system of two integro-differential equations which form a new representation of the solution to the inviscid BE. This formalism clarifies the relation between pole dynamics and branch-cut dynamics. A large wave number asymptotic expansion of the Fourier transform of the inviscid solution uniformly valid in a neighborhood of the shock time is described in terms of the Airy function. This provides a clarification of the degeneracy presented by Fournier and Frisch in [19]. In [33], this methodology is adapted to the dispersive case $\nu \in i\mathbb{R}$.

2. Integral representation, pole expansion, and pole dynamics for $\nu > 0$.

2.1. The Cole–Hopf solution and Mittag–Leffler expansion.

The Cole–Hopf solution to the initial value problem (1.1)–(1.2) can be represented by a Mittag–Leffler expansion as follows.

THEOREM 2.1. *For all $\nu, t, t_* > 0$, the solution to BE with initial data $u_0(x) = 4x^3 - x/t_*$ is*

$$u_\nu(x,t) = \frac{x}{t} - 2\nu\,\partial_x \log\bigl(E_\nu(x,t)\bigr),$$
$$E_\nu(x,t) = \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\nu}\left(\frac{x}{t}y + \alpha y^2 - y^4\right)\right\} dy,$$

where $2\alpha = 1/t_* - 1/t \in \mathbb{R}$. For fixed ν, t, $E_\nu(x,t)$ is an even entire function of x of order $4/3$ with countably many simple zeros which come in pure imaginary opposite *and conjugate pairs. Moreover, $E_\nu(x,t)$ has the infinite product representation*

$$E_\nu(x,t) = \mathcal{C}_\nu(t) \prod_{n=1}^{\infty} \left(1 + \frac{x^2}{\beta_n^2(t,\nu)}\right), \quad \sum_{n=1}^{\infty} \frac{1}{\beta_n} = +\infty, \quad \sum_{n=1}^{\infty} \frac{1}{\beta_n^2} < +\infty,$$
$$\mathcal{C}_\nu(t) = \frac{\sqrt{\alpha}}{2}\, e^{\frac{\alpha^2}{16\nu}} K_{1/4}\left(\frac{\alpha^2}{16\nu}\right), \quad \mathcal{C}_\nu(t_*) = \nu^{1/4} 2^{-3/4} \Gamma(1/4),$$

where $K_q(z)$ is the modified Bessel function of the second kind. Thus, the solution $u_\nu(x,t)$ has an alternate representation in terms of a Mittag–Leffler (pole) expansion

$$u_\nu(x,t) = \frac{x}{t} - \sum_{n=1}^{\infty} \frac{4\nu x}{x^2 + \beta_n^2(t,\nu)} = \frac{x}{t} - 2\nu \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{x - i\beta_n(t,\nu)},$$

which converges uniformly on compact sets for x away from the poles $a_n = \pm i\beta_n$.

Proof. The solution to system (1.1) is constructed using the Cole–Hopf nonlinear transform $u = -2\nu\, \partial_x \log(\phi_\nu)$ [17, 22], which was first introduced by Forsyth (cf. [18, section 207, p. 100]). This nonlinear dependent variable transformation maps BE into the diffusion equation for $\phi_\nu(x,t)$ with corresponding initial data $\phi_0(x) = \exp\{-\frac{1}{2\nu}\int_0^x u_0(y)dy\}$. The solution is therefore represented by means of a convolution:

$$\begin{aligned} \phi_\nu(x,t) &= (K_\nu * \phi_o)(x,t) \\ &= (4\pi\nu t)^{-1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-y)^2}{4\nu t} - \frac{1}{2\nu}\int_0^y u_0(\eta)d\eta\right\} dy \\ &= K_\nu(x,t) E_\nu(x,t), \end{aligned}$$

where

$$K_\nu(x,t) = K(x,\nu t) = (4\pi\nu t)^{-1/2} \exp\left(-x^2/4\nu t\right)$$

is the fundamental solution of the diffusion equation and

$$E_\nu(x,t) = \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\nu}\int_0^y \left(\frac{x}{t} - \frac{\eta}{t} - u_0(\eta)\right) d\eta\right\} dy. \tag{2.1a}$$

Since $\partial_x \log\big(K_\nu(x,t)\big) = -x/2\nu t$, the solution of the original problem is given by

$$u_\nu(x,t) = \frac{x}{t} - 2\nu\, \partial_x \log\big(E_\nu(x,t)\big). \tag{2.1b}$$

For $u_0(x) = 4x^3 - x/t_*$, $\nu, t > 0$, we obtain the following solution:

$$E_\nu(x,t) = \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\nu}\left(\frac{x}{t}y + \alpha y^2 - y^4\right)\right\} dy, \qquad \alpha = \frac{t - t_*}{2tt_*} \in \mathbb{R}. \tag{2.2}$$

It is clear that E_ν is an even, real analytic function of x and therefore satisfies the conjugacy relation $E_\nu(\overline{x},t) = \overline{E_\nu(x,t)}$ (the analyticity of E_ν can be verified using Morera's theorem). The positive order λ of an entire function $f(z)$ is defined as $\lambda = \limsup_{r\to+\infty} \log\log M(r)/\log r$, where $M(r) = \max_{|z|=r}|f(z)|$. For a fixed time $t > 0$, the order of E_ν is the smallest number $\lambda \in \mathbb{R}_+$ such that

$M_\nu(r) = \max_{|x|=r} |E_\nu(x,t)| \leq \exp(r^{\lambda+\epsilon})$ for any $\epsilon > 0$ as soon as r is sufficiently large. From the asymptotic behavior of E_ν for $|x| = r \to +\infty$, we find in (4.4) that $M_\nu(r) = \mathcal{O}(r^{1/3}\exp(-\kappa(t)\, r^{4/3}/2\nu))$, where $\kappa(t)/2\nu$ is the "type" of the entire function E_ν. Thus, it is clear that its order is $\lambda = 4/3$. It is known that entire functions of fractional order have infinitely many zeros (see [3, 8]); thus, E_ν has infinitely many zeros that come in opposite and conjugate pairs. Since the fractional order of the entire function E_ν is also the exponent of convergence of its zeros a_n (see again [3, 8]), we have

$$\sum_{n=1}^{\infty} \frac{1}{|a_n|^{\lambda+\epsilon}} < +\infty \ \forall \epsilon > 0. \tag{2.3}$$

Using a Hadamard decomposition, we construct the solution u_ν by factorization of the zeros of E_ν. The canonical infinite product expansion of E_ν is (see [3])

$$E_\nu(x,t) = \mathcal{C}x^m e^{g(x)} \prod_{n=1}^{\infty} \left(1 - \frac{x}{a_n}\right) e^{x/a_n + \frac{1}{2}(x/a_n)^2 + \cdots + \frac{1}{p}(x/a_n)^p},$$

where $g(x)$ is a polynomial of degree q. The integer $h = \max(p,q)$, which is called the genus of the product representation of the entire function E_ν, satisfies the bound $h \leq \lambda \leq h+1 \Rightarrow h = 1 \Rightarrow p,q \leq 1$. Moreover, since E_ν is an even function of x, we must have $q = \deg g(x) = 0$, and therefore $p = 1$ (since $p+1 > \lambda$, $p \in \mathbb{N}$). Since $\mathcal{C} = \mathcal{C}_\nu(t) = E_\nu(0,t) \neq 0$ (see (2.6a)), we must also set $m = 0$, so the canonical product must be of the form

$$E_\nu(x,t) = \mathcal{C}_\nu(t) \prod_{n=1}^{\infty} \left(1 - \frac{x}{a_n}\right) e^{x/a_n}, \quad \sum_{n=1}^{\infty} \frac{1}{|a_n|} = +\infty, \quad \sum_{n=1}^{\infty} \frac{1}{|a_n|^2} < +\infty.$$

Due to the even parity of E_ν, its zeros come in opposite pairs $x = \pm a_n$; thus, the product representation reduces to the simple form

$$E_\nu(x,t) = \mathcal{C}_\nu(t) \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{a_n^2}\right).$$

In [29], Pólya showed that functions of the form

$$\int_{-\infty}^{\infty} e^{-at^{4n} + bt^{2n} + iyt} dt \qquad n \geq 1,\ a > 0,\ b \in \mathbb{R} \tag{2.4}$$

have only real zeros. Using this property, it is straightforward to show that the zeros of E_ν come in pure imaginary conjugate pairs; thus, we let $a_n = i\beta_n$, $\beta_n > 0$ and obtain an infinite product expansion of E_ν valid for all $t, \nu > 0$:

$$E_\nu(x,t) = \mathcal{C}_\nu(t) \prod_{n=1}^{\infty} \left(1 + \frac{x^2}{\beta_n^2(t,\nu)}\right), \quad \sum_{n=1}^{\infty} \frac{1}{\beta_n} = +\infty, \quad \sum_{n=1}^{\infty} \frac{1}{\beta_n^2} < +\infty, \tag{2.5}$$

where $\mathcal{C}_\nu(t)$ is a constant depending on t which can be found explicitly: let $K_q(z)$ be the modified Bessel function of the second kind; then

$$\mathcal{C}_\nu(t) = E_\nu(0,t) = \int_{-\infty}^{\infty} e^{(\alpha y^2 - y^4)/2\nu}\, dy = \frac{\sqrt{\alpha}}{2}\, e^{\frac{\alpha^2}{16\nu}}\, K_{1/4}\left(\frac{\alpha^2}{16\nu}\right), \tag{2.6a}$$

$$\mathcal{C}_\nu(t_*) = E_\nu(0,t_*) = \int_{-\infty}^{\infty} e^{-y^4/2\nu}\, dy = \nu^{1/4} 2^{-3/4}\Gamma(1/4), \tag{2.6b}$$

with $K_{1/4}(z) = \mathcal{O}(z^{-1/4})$ as $z \to 0$. After logarithmic differentiation of E_ν and by using (2.1b) and (2.5), the spatially singular part of the solution being the ratio of two entire functions is meromorphic. Thus, we obtain a Mittag–Leffler expansion of the solution which we refer to as the pole expansion:

$$u_\nu(x,t) = \frac{x}{t} - \sum_{n=1}^{\infty} \frac{4\nu x}{x^2 + \beta_n^2(t,\nu)} = \frac{x}{t} - 2\nu \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{x - i\beta_n(t,\nu)}. \tag{2.7}$$

In the second sum, we have adopted the convention that $\beta_{-n} = -\beta_n$. Furthermore, it must be understood as a symmetric (convergent) sum of the form

$$\sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{x - i\beta_n(t,\nu)} = \sum_{n\in\mathbb{N}^*} \left(\frac{1}{x - i\beta_n(t,\nu)} + \frac{1}{x + i\beta_n(t,\nu)} \right).$$

Since $\sum_n \beta_n^{-2} < \infty$ for any fixed $t, \nu > 0$, the series defining u_ν in (2.7) converges absolutely and uniformly on any strip $0 < \beta_k < \delta_k \leq |\Im x| \leq \delta_{k+1} < \beta_{k+1}$, $k \in \mathbb{N}^* = \mathbb{N}\backslash\{0\}$. Therefore, u_ν is analytic in the strip $|\Im x| < \beta_1$ where $i\beta_1$ is the first ordered pole on the imaginary axis. From (2.7), u_ν conserves the odd parity of the initial data as expected from the PDE: $u_\nu(-x,t) = -u_\nu(x,t)$. In order for this pole expansion to make sense, the behavior of the spatially singular part of the expansion should be unbounded as $t \to 0^+$ in order to balance with the term x/t. □

2.2. CDS for the poles $\boldsymbol{\beta_n(t,\nu)}$. We describe the time evolution of the poles $\beta_n(t,\nu)$ according to an infinite dimensional dynamical system which is found as a compatibility condition for the existence of the pole expansion (2.7). We prove the following property.

Property 2.2. *The imaginary part $\beta_n = \beta_n(t,\nu) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ of the simple poles $x = \pm i\beta_n$ of $u_\nu(x,t)$ satisfy the Calogero-type infinite-dimensional dynamical system*

$$\dot{\beta}_n = \frac{\beta_n}{t} + \frac{\nu}{\beta_n} - 4\nu\beta_n \sum_{\substack{l=1 \\ l\neq n}}^{\infty} \frac{1}{\beta_l^2 - \beta_n^2} \ \forall n \in \mathbb{N}^*.$$

Adopting the convention that $\beta_{-n} = -\beta_n$, we have a symmetric formulation:

$$\dot{\beta}_n = \frac{\beta_n}{t} - 2\nu \sum_{\substack{l=-\infty \\ l\neq n,0}}^{\infty} \frac{1}{\beta_l - \beta_n} \ \forall n \in \mathbb{Z}^*.$$

Moreover, the variables $\gamma_n(t,\nu) = \beta_n^2(t,\nu)/\nu$ satisfying $\sum_n \gamma_n^{-1} < +\infty$ are the solution to the ν-independent infinite system of ODEs:

$$\frac{\dot{\gamma}_n}{2} = \frac{\gamma_n}{t} + 1 - 4\gamma_n \sum_{\substack{l=1 \\ l\neq n}}^{\infty} \frac{1}{\gamma_l - \gamma_n} \ \forall n \in \mathbb{N}^*.$$

Proof. The usual pole expansion that is sought in [1, pp. 203–209] and [13, 16, 20] is of the form

$$u_\nu = \sum_{n=1}^{N} \frac{2\nu}{x - i\beta_n}. \tag{2.8}$$

However, as $N \to +\infty$ this series diverges for any fixed x, t, ν. Since we know that the representation (2.7) converges away from the poles (since $\sum_n \beta_n^{-2} < \infty$ from (2.5)) instead of using (2.8), we replace the full Mittag–Leffler/pole expansion (2.7) in the PDE (1.1). The extension from a finite pole expansion to an infinite one is easily made (cf. [16, section 3]). We obtain an infinite system of ODEs which govern the motion of the poles $\beta_n(t,\nu)$ as they evolve with time for $t > 0$. We introduce the following notations:

$$\dot{\beta}_n = \frac{d\beta_n}{dt}, \qquad \sum_n = \sum_{n=1}^{\infty}, \qquad \sum_l = \sum_{l=1}^{\infty}, \qquad \sum_{l\neq n} = \sum_{\substack{l=1 \\ l\neq n}}^{\infty}.$$

Using partial fraction expansion we have the following property.

PROPERTY 2.3. *Let $x \notin \{i\beta_n \in i\mathbb{R}$ for all $n \in \mathbb{Z}^*\}$ and $l \neq n$; then*

$$\text{(2.9a)} \qquad \frac{1}{(x^2+\beta_n^2)(x^2+\beta_l^2)} = \frac{1}{\beta_n^2-\beta_l^2}\cdot\left(\frac{1}{x^2+\beta_l^2} - \frac{1}{x^2+\beta_n^2}\right)$$

$$\text{(2.9b)} \qquad \begin{aligned} \frac{1}{(x^2+\beta_n^2)(x^2+\beta_l^2)^2} = \frac{1}{(\beta_l^2-\beta_n^2)^2}&\cdot\left(\frac{1}{x^2+\beta_n^2} - \frac{1}{x^2+\beta_l^2}\right) \\ &+ \frac{1}{\beta_n^2-\beta_l^2}\cdot\frac{1}{(x^2+\beta_l^2)^2}. \end{aligned}$$

Due to obvious symmetries, we also have

$$\text{(2.10)} \qquad \sum_n\sum_{l\neq n}\frac{1}{(\beta_n^2-\beta_l^2)^2}\cdot\left(\frac{1}{x^2+\beta_n^2} - \frac{1}{x^2+\beta_l^2}\right) = 0.$$

Thus, combining (2.9b) and (2.10), we obtain Property 2.4.

PROPERTY 2.4. *Let $x \notin \{i\beta_n \in i\mathbb{R}$ for all $n \in \mathbb{Z}^*\}$; then*

$$\begin{aligned} \sum_n\sum_{l\neq n}\frac{1}{(x^2+\beta_n^2)}\cdot\frac{1}{(x^2+\beta_l^2)^2} &= \sum_n\sum_{l\neq n}\frac{1}{\beta_n^2-\beta_l^2}\cdot\frac{1}{(x^2+\beta_l^2)^2} \\ &= \sum_n\sum_{l\neq n}\frac{1}{\beta_l^2-\beta_n^2}\cdot\frac{1}{(x^2+\beta_n^2)^2}. \end{aligned}$$

After replacing the pole expansion (2.7) into the original PDE (1.1), canceling the terms $\pm x/t^2$, regrouping terms together in powers of $(x^2+\beta_n^2)^{-1}$, dividing by $8\nu x$, and appealing to both equation (2.9a) of Properties 2.3 and 2.4, Property 2.2 is proved. Note that in [7], the pole interaction in the dynamical system was incorrectly stated as a divergent semi-infinite sum of the form

$$\dot{\beta}_n = \frac{\beta_n}{t} - 2\nu\sum_{\substack{l\geq 1 \\ l\neq n}}\frac{1}{\beta_l-\beta_n}.$$

Due to the generality of the integral representation (2.1a,b) of solutions to an initial value problem (IVP) for BE, it is important to see that the pole representation

(2.7) and pole dynamics of Property 2.2 are special cases of the general representation for a meromorphic solution to BE.

PROPERTY 2.5. *Let $\mathcal{I} \subseteq \mathbb{Z}$ be a set of indices, either finite or countable, and let $\{a_n(t,\nu), n \in \mathcal{I}\}$ be a finite or countable set of time-dependent pole locations. Then any meromorphic solution with poles $\{a_n(t,\nu), n \in \mathcal{I}\}$ must have the following pole representation/pole dynamics:*

$$\textit{Pole representation:}\quad u_\nu(x,t) = \frac{x}{t} - 2\nu \sum_{n\in\mathcal{I}} \frac{1}{x - a_n(t,\nu)},$$

$$\textit{Pole dynamics:}\quad \dot{a}_n = \frac{a_n}{t} - 2\nu \sum_{\substack{l\in\mathcal{I}\\ l\neq n}} \frac{1}{a_n - a_l} \qquad \forall n \in \mathcal{I}.$$

Moreover, if the initial data is odd, then the poles must come in opposite pairs $\{a_{\pm n}(t,\nu), n \in \mathcal{I} \subseteq \mathbb{Z}^* = \mathbb{Z}\backslash\{0\} \mid a_{-n} = -a_n\}$. In this case, the pole representation is fully symmetric.

3. Asymptotic analysis of $u_\nu(x,t)$ as $\nu \to 0^+$, $t > t_*$.

When $\nu \to 0^+$, we evaluate the asymptotic behavior of E_ν using the saddle-point method. The caustic $x = x_s(t)$ corresponds to the envelope of the characteristics of the inviscid Burgers solution and is also determined by the following system of equations:

$$\left\{ \begin{array}{l} 0 = w_z(z,x) = x/t + 2\alpha z - 4z^3, \\ 0 = w_{zz}(z,x) = 2\alpha - 12z^2, \end{array} \right. \tag{3.1}$$

where

$$w(z,x) = \int_0^z \left(\frac{x}{t} - \frac{\eta}{t} - u_0(\eta) \right) d\eta \tag{3.2}$$

is the phase function of the integrand in the definition of $E_\nu(x,t)$. This system represents the conditions for the phase function w to have saddle points of multiplicity two, thereby yielding a curve in the (x,t) plane on which two saddle points of multiplicity one coalesce into a saddle point of multiplicity two. From the second equation in (3.1), we find $z_{caustic}(t) = \pm\sqrt{\alpha/6}$; from the first, we have

$$x = x_{caustic} = t\big(4z_{caustic}(t)^3 - 2\alpha z_{caustic}(t)\big) = \mp t \left(\frac{2\alpha}{3}\right)^{3/2} = \mp x_s(t), \tag{3.3}$$

where $x_s(t) = (3t_*)^{-3/2}(t-t_*)^{3/2}t^{-1/2}$ is the second-order branch point of the inviscid solution described in [32, App. C]. We find that all three saddle points may be relevant within the caustic $|x| < |x_s(t)| - \delta/2$, where $\delta > 0$. For a discussion on such caustics, cf. [24, 28]. When $t > t_*$, $x \in \big(-\infty, -x_s(t) - \delta/2\big) \cup \big(x_s(t) + \delta/2, \infty\big)$, $\nu \to 0^+$, the same analysis holds and one recovers the characteristic solution outside the caustic consisting of only one relevant saddle point. The transition from within the caustic to outside is not uniform as the asymptotic behavior at the caustic $x = \pm x_s(t)$ is degenerate (two saddle points have coalesced). The transitionary regime from one relevant saddle point to two at and around the caustic is therefore described by means of the Pearcey integral which allows for a uniformly valid description.

3.1. Inner expansion: $x \in \big(-x_s(t)+\delta/2,\ x_s(t)-\delta/2\big)$, $\delta > 0$, $t > t_*$. In the analysis that follows, we are only concerned with the dominant behavior of E_ν; thus, we only retain the first term:

$$E_\nu(x,t) = \sum_{s=0,1,2} \sqrt{\frac{-4\pi\nu}{w_{zz}(z_s,x)}} \exp\left(\frac{w(z_s,x)}{2\nu}\right)\Big(1+\mathcal{O}(\nu)\Big), \tag{3.4}$$

as $\nu \to 0^+$, with

$$w_z\big(z_s(x,t),x\big) = 0, \quad w_{zz}\big(z_s(x,t),x\big) = 2\alpha - 12z_s^2. \tag{3.5}$$

Since

$$0 = \frac{z_s}{4} w_z\big(z_s,x\big) = \frac{xz_s}{4t} + \frac{\alpha}{2}z_s^2 - z_s^4,$$

we have that

$$w\big(z_s(x,t),x\big) = \frac{x}{t}z_s + \alpha z_s^2 - z_s^4 = \frac{3}{4}\frac{x}{t}z_s + \frac{\alpha}{2}z_s^2. \tag{3.6}$$

The values of the saddle points $z_s = z_s(x,t)$ of (3.4) are determined by the three roots of the first equation in system (3.1), i.e., the first equation of (3.5). They are, specifically,

$$\begin{cases} z_0 = \omega\,\mathcal{A} + \omega^2\,\mathcal{B}, \\ z_1 = \omega^2\,\mathcal{A} + \omega\,\mathcal{B}, \\ z_2 = \mathcal{A} + \mathcal{B}, \end{cases} \tag{3.7}$$

where $w = e^{2\pi i/3}$ is a cube root of unity and

$$\begin{cases} \mathcal{A}(x,t) = (8t)^{-1/3} \cdot \sqrt[3]{x + \sqrt{x^2 - x_s^2}}, \\ \mathcal{B}(x,t) = (8t)^{-1/3} \cdot \sqrt[3]{x - \sqrt{x^2 - x_s^2}}. \end{cases} \tag{3.8}$$

Note that all three saddle points are real when $x, x_s \in \mathbb{R}$ and the discriminant $\Delta = x^2 - x_s^2 < 0$, that is, $|x| < |x_s(t)|$. In this case, $\mathcal{A} = \overline{\mathcal{B}}$ (see [32, App. B]). Therefore, we have $z_s \in \mathbb{R}$, $w(z_s,x) \in \mathbb{R}$, and $w_{zz}(z_s,x) = 2\alpha - 12z_s^2 \in \mathbb{R}$. Hence, all three terms in the summation signs may be relevant. Note, however, that the expansion derived for E_ν is only valid within $|x| < |x_s|$, and in order to get an expansion uniformly valid across $x = \pm x_s$ one needs to derive a uniform expansion as presented in section 3.3. The dominant behavior of the solution $u_\nu(x,t)$ is found from the Cole–Hopf representation, so within the caustic $|x| < |x_s| - \delta/2$ we have

$$\begin{aligned} \frac{U_\nu(x,t)}{t} &= 2\nu\,\partial_x \log\big(E_\nu(x,t)\big) \\ &= 2\nu\,\partial_x \log\left(\sum_{s=0,1,2}\sqrt{\frac{-4\pi\nu}{w_{zz}(z_s,x)}}e^{\frac{w(z_s,x)}{2\nu}}\big(1+\mathcal{O}(\nu)\big)\right) \\ &= 2\nu\,\frac{\sum_{s=0,1,2}\partial_x\left(\sqrt{\frac{-4\pi\nu}{w_{zz}(z_s,x)}}e^{\frac{w(z_s,x)}{2\nu}}\right)}{\sum_{s=0,1,2}\sqrt{\frac{-4\pi\nu}{w_{zz}(z_s,x)}}e^{\frac{w(z_s,x)}{2\nu}}} + \mathcal{O}(\nu^2). \end{aligned}$$

Since $w(z_s,x)\in\mathbb{R}$, $\nu>0$, and

$$\frac{\partial w}{\partial x}(z_s,x)=\frac{z_s}{t}, \tag{3.9}$$

we find

$$U_\nu(x,t)=\frac{\sum_{s=0,1,2} z_s\cdot e^{\frac{w(z_s,x)}{2\nu}}/\sqrt{w_{zz}(z_s,x)}}{\sum_{s=0,1,2} e^{\frac{w(z_s,x)}{2\nu}}/\sqrt{w_{zz}(z_s,x)}}+\mathcal{O}(\nu). \tag{3.10}$$

The x-differentiation of the asymptotic formula of $E_\nu(x,t)$ is justified due to the analytic dependency in x. Often one of the three saddle points is such that $w(z_s,x)<0$, and as such, its contribution is exponentially smaller than either of the other two. In terms of the numerical computation that will be carried out in section 5, leaving this term in (3.1) does not affect the value of u_ν. Thus, we can simplify expression (3.10) to a two-term asymptotic expansion that is similar to the one in [38, section 4.2]. Clearly, the further away we are from the caustic, the more dominant one of the saddle points becomes. However, since there is a point where the dominance of one over the other changes (i.e., where they are equally relevant), we must leave both in the asymptotic formula. Note also that $w_{zz}(z_s,x)\to\infty$ as $x\to x_s$, which is characteristic of the degeneracy of the asymptotic formula (3.10) at the caustic $x=x_s$.

PROPERTY 3.1. *For $x\in\big(-x_s(t)+\delta/2,x_s(t)-\delta/2\big)$, $\delta>0$, $t>t_*$, the inner expansion of the solution to BE as $\nu\to0^+$ is given by*

$$u_\nu(x,t)=\frac{x}{t}-\frac{U_\nu(x,t)}{t},$$
$$U_\nu(x,t)=\frac{\sum_{\{s:w(z_s,x)>0\}} z_s\cdot e^{\frac{w(z_s,x)}{2\nu}}/\sqrt{w_{zz}(z_s,x)}}{\sum_{\{s:w(z_s,x)>0\}} e^{\frac{w(z_s,x)}{2\nu}}/\sqrt{w_{zz}(z_s,x)}}+\mathcal{O}(\nu).$$

3.2. Outer expansion: $x\in\big(-x_s(t)-\delta/2,x_s(t)+\delta/2\big)^c$, $\delta>0$, $t>t_*$.

The inviscid limit is found in a straightforward manner in this case: only one saddle point is relevant, so the asymptotic limit derived in section 3.1 reduces to

$$U_\nu(x,t)=U(x,t)+\mathcal{O}(\nu)\qquad\text{as }\nu\to0^+,$$

where $U(x,t)=z_{s*}(x,t)$ is the spatially singular part of the inviscid solution (see [32, App. C]). The particular saddle point z_{s*} that is chosen at every x is the one for which $w(z_{s*},x)=\max_{s=0,1,2}w(z_s,x)$. Hence, we have the following property outside of the caustic.

PROPERTY 3.2. *Let $\delta>0$, $t>t_*$, and define $z_{s*}(x,t)$ by*

$$w(z_{s*},x)=\max_{s=0,1,2}w(z_s,x).$$

Then for $x\in\big(-x_s(t)-\delta/2,x_s(t)+\delta/2\big)^c$, the solution to BE is given by

$$u_\nu(x,t)=\frac{x}{t}-\frac{U_\nu(x,t)}{t}=\frac{x}{t}-\frac{U(x,t)}{t}+\mathcal{O}(\nu)\qquad\textit{as }\nu\to0^+,$$

where $U(x,t)=z_{s*}(x,t)$ is the Lagrangian characteristic variable of the inviscid solution.

3.3. Uniform asymptotic expansion as $\nu \to 0^+$ in a neighborhood of the caustics $x = \pm x_s(t)$ for $t > t_*$ via Pearcey's integral. Following the notation of Kaminski in [24], we introduce the Pearcey integral from which one can derive a uniform asymptotic expansion with two coalescing saddle points (see [15]): let

$$P(X,Y) = \int_{-\infty}^{+\infty} e^{i\left(v^4/4 + Xv^2/2 + Yv\right)}\,dv \tag{3.11}$$

denote the Pearcey integral. Introducing the change of variable

$$y \to (-i\nu/2)^{1/4}\,v = (\nu/2)^{1/4}\,e^{3\pi i/8}\,v$$

and deforming the path of integration back to the real axis using Jordan's lemma, we can express $E_\nu(x,t)$ as

$$\begin{aligned} E_\nu(x,t) &= \int_{-\infty}^{+\infty} \exp\left\{\frac{1}{2\nu}\left(\frac{x}{t}y + \alpha y^2 - y^4\right)\right\} dy \\ &= \left(\frac{-i\nu}{2}\right)^{1/4} \int_{-\infty}^{+\infty} \exp\left\{ i\left(\frac{v^4}{4} + \frac{\alpha e^{i\pi/4}}{\sqrt{2\nu}}\frac{v^2}{2} + \frac{xe^{-i\pi/8}}{2t}\left(\frac{1}{2\nu^3}\right)^{1/4} v\right)\right\} dv \\ &= \left(\frac{-i\nu}{2}\right)^{1/4} P\left(X = \frac{\alpha e^{i\pi/4}}{\sqrt{2\nu}}, Y = \frac{xe^{-i\pi/8}}{2t}\left(\frac{1}{2\nu^3}\right)^{1/4}\right). \end{aligned} \tag{3.12}$$

Clearly, a small ν asymptotic of E_ν is equivalent to a combined asymptotic expansion of the Pearcey integral as $|X|, |Y| \to +\infty$. The caustic of $P(X,Y)$ and the corresponding caustic of $E_\nu(x,t)$ is given by

$$Y = \frac{2}{\sqrt{27}}(-X)^{3/2} \Longleftrightarrow x = \pm x_s(t) \in \mathbb{R} \quad \text{for } t > t_*. \tag{3.13}$$

Hence, the uniform asymptotic behavior of E_ν in a neighborhood of the caustic is found from the one of $P(-X, (2/\sqrt{27} - \tau)X^{3/2})$ as $X \to +\infty$, where $\tau = 0$ at the caustic and $\tau \neq 0$ away from it (see [24]). This amounts to a uniformly valid expansion in the interval $|x \pm x_s(t)| \leq |\delta_\pm(\tau;t)|$, where $\delta_\pm(\tau) = \delta_\pm(\tau;t) = \mp\sqrt{27}x_s(t)\cdot\tau/2 \in \mathbb{R}$. This expansion is also valid outside of this interval, however the region of interest is a neighborhood of the caustic. Indeed one only needs to use the asymptotic expansion of the Airy function and its derivative to find the results obtained in sections 3.1 and 3.2. From (3.12) we have that

$$\begin{aligned} U_\nu(x,t) &= t\cdot 2\nu\,\partial_x \log\bigl(E_\nu(x,t)\bigr) \\ &= t\cdot 2\nu\,\partial_x \log\left[P\left(X(\nu;t) = \frac{\alpha e^{i\pi/4}}{\sqrt{2\nu}}, Y(\nu;x,t) = \frac{xe^{-i\pi/8}}{2t}\left(\frac{1}{2\nu^3}\right)^{1/4}\right)\right]. \end{aligned}$$

Let

$$X = X(\nu;t), \qquad Y = Y(\nu;x,t) = Y(\nu; x = \pm x_s(t) - \delta_\pm(\tau;t), t),$$

where $\delta_\pm(\tau;t) \to 0^\mp$ as $\tau \to 0^\pm$, so that

$$\begin{aligned} U_\nu\Bigl(x = \pm x_s(t) - \delta_\pm(\tau;t), t\Bigr) &= t\cdot 2\nu\,\partial_x \log P(X,Y) \\ &= -t\cdot 2\nu\,\partial_\tau \log\Bigl(P\left(-X, (2/\sqrt{27} - \tau)X^{3/2}\right)\Bigr) \Big/ \frac{\partial \delta_\pm}{\partial\tau}\,. \end{aligned}$$

Let $P(\tau) = P\left(-X, (2/\sqrt{27} - \tau)X^{3/2}\right)$. Then, since $\partial\delta_\pm/\partial\tau = \mp\sqrt{27}x_s(t)/2$, we have

$$U_\nu\Big(x = \pm x_s(t) - \delta_\pm(\tau; t), t\Big) = \pm\frac{4\nu t}{\sqrt{27}x_s(t)}\frac{P_\tau(\tau)}{P(\tau)}.$$

Let

$$Ai(z) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{v^3/3 - zv}\, dv$$

stand for the Airy function (cf. [2]); then the following property is found in [24].

Property 3.3. *The uniform asymptotic expansion of* $P(-X, (\frac{2}{\sqrt{27}} - \tau)X^{3/2})$ *as* $X \to +\infty$ *in a neighborhood of* $\tau = 0$ *is given by*

$$P\left(-X, \left(\frac{2}{\sqrt{27}} - \tau\right)X^{3/2}\right) = \Bigg[e^{iX^2[f(v_2)+f(v_3)]/2}\Bigg\{p_0(\tau)\frac{2\pi}{X^{1/6}}Ai(-X^{4/3}\zeta(\tau))$$
$$+q_0(\tau)\frac{2\pi}{iX^{5/6}}Ai'(-X^{4/3}\zeta(\tau))\Bigg\} + e^{iX^2 f(v_1)}\left(\frac{\pi}{3v_1^2 - 1}\right)^{1/2}\frac{1+i}{X^{1/2}}\Bigg]\left(1 + \mathcal{O}\left(\frac{1}{X^2}\right)\right),$$

with

$$p_0(\tau) = 3^{-1/6}(1 + \mathcal{O}(\tau)), \quad q_0(\tau) = -\frac{3^{-5/6}}{2}(1 + \mathcal{O}(\tau)), \quad \zeta(\tau) = 3^{-1/6}\tau(1 + \mathcal{O}(\tau)),$$

and

$$f(v) = f(v;\tau) = \frac{v^4}{4} - \frac{v^2}{2} + \left(\frac{2}{\sqrt{27}} - \tau\right)v,$$

and the $v_i, i = 1, 2, 3$ *are the saddle points of* $f(v;\tau)$ *determined by the equation* $f_v(v_i;\tau) = 0$*, so* $f(v_i;\tau) = -v_i^2/4 + (2/\sqrt{27} - \tau)3v_i/4$*. The* v_i*'s are, specifically,*

$$v_1(\tau) = -\frac{2}{\sqrt{3}}\sin\left(\frac{\pi}{3} + \phi(\tau)\right), v_2(\tau) = \frac{2}{\sqrt{3}}\sin(\phi(\tau)), v_3(\tau) = \frac{2}{\sqrt{3}}\sin\left(\frac{\pi}{3} - \phi(\tau)\right),$$

where

$$\phi = \phi(\tau) = \frac{1}{3}\arcsin\left(1 - \tau\sqrt{27}/2\right), \quad \tau \in \mathbb{R}, \quad |\phi| \le \frac{\pi}{6}.$$

In order to derive the uniform asymptotic expansion of the derivative P_τ, one can differentiate termwise the expression in Property 3.3 due to the analytic dependency of $P(X, Y)$ in both its arguments X, Y (see [24] and [37, p. 52]). Therefore, since

$$X = \frac{\alpha e^{i\pi/4}}{\sqrt{2\nu}} \Rightarrow \frac{iX^2}{2} = -\frac{\alpha^2}{4\nu} \Rightarrow X^{-2} = \mathcal{O}(\nu)$$

and

$$\frac{\partial f}{\partial\tau}(v_i;\tau) = -v_i, \quad \frac{\partial f}{\partial v}(v_i;\tau) = 0 \quad \Rightarrow \quad \frac{df}{d\tau}(v_i(\tau);\tau) = -v_i(\tau),$$

and using the fact that $2\alpha/3 = (x_s/t)^{2/3}$, the next property is proved.

Property 3.4. *Let $\delta_\pm(\tau;t) = \mp\sqrt{27}x_s(t)\cdot\tau/2$. Then the uniform asymptotic expansion as $\nu \to 0^+$ of $U_\nu(x = \pm x_s(t) - \delta_\pm(\tau;t), t)$ in a neighborhood of the caustic $x = \pm x_s(t)$ is*

$$\begin{aligned}U_\nu\Big(x = \pm x_s(t) - \delta_\pm(\tau;t), t\Big) = \pm\,\frac{\sqrt{3}}{2}\left(\frac{x_s(t)}{t}\right)^{1/3} &\times \Bigg[[v_2+v_3]e^{-\frac{\alpha^2}{4\nu}[f(v_2)+f(v_3)]}\\ \times\left\{p_0(\tau)\frac{2\pi}{X^{1/6}}Ai(-X^{4/3}\zeta(\tau)) + q_0(\tau)\frac{2\pi}{iX^{5/6}}Ai'(-X^{4/3}\zeta(\tau))\right\}&\\ +2v_1\,e^{-\frac{\alpha^2}{4\nu}2f(v_1)}\left(\frac{\pi}{3v_1^2-1}\right)^{1/2}\frac{1+i}{X^{1/2}}\Bigg]&\\ \Bigg/\Bigg[e^{-\frac{\alpha^2}{4\nu}[f(v_2)+f(v_3)]}\left\{p_0(\tau)\frac{2\pi}{X^{1/6}}Ai(-X^{4/3}\zeta(\tau)) + q_0(\tau)\frac{2\pi}{iX^{5/6}}Ai'(-X^{4/3}\zeta(\tau))\right\}&\\ +e^{-\frac{\alpha^2}{4\nu}2f(v_1)}\left(\frac{\pi}{3v_1^2-1}\right)^{1/2}\frac{1+i}{X^{1/2}}\Bigg] + \mathcal{O}(\nu) \qquad \textit{as } \nu \to 0^+.&\end{aligned}$$

3.3.1. Behavior at the caustics $x = \pm x_s(t)$.

At the caustics $x = \pm x_s(t)$, $\tau = 0$, $\phi(0) = \pi/6$, $v_1(0) = -2/\sqrt{3}$, and $v_2(0) = v_3(0) = 1/\sqrt{3}$. Moreover, $f(v_i;0) = -v_i^2/4 + v_i/2\sqrt{3}$, so $f(v_2;0) = f(v_3;0) = -2/3$ and $f(v_1;0) = 1/12$. Since $f(v_2;0) < 0$ and $f(v_1;0) > 0$, the dominant term as $\nu \to 0^+$ in both the numerator and denominator of U_ν is obviously the one containing the exponentially increasing factor $\exp(-\frac{\alpha^2}{4\nu}[f(v_2) + f(v_3)])$. Therefore, the dominant behavior of $U_\nu(\pm x_s(t), t)$ reduces to the simple form

$$\begin{aligned}U_\nu(\pm x_s(t), t) &= \pm\frac{\sqrt{3}}{2}\left(\frac{x_s(t)}{t}\right)^{1/3}\cdot(v_2(0) + v_3(0)) + \mathcal{O}(\nu)\\ &= \pm\left(\frac{x_s(t)}{t}\right)^{1/3} + \mathcal{O}(\nu) \quad \text{as } \nu \to 0^+.\end{aligned}$$

Thus, since $u_\nu(x,t) = x/t - U_\nu(x,t)/t$ and from the odd parity of u_ν, we have the following property.

Property 3.5. *The asymptotic behavior of the solution $u_\nu(x,t)$ as $\nu \to 0^+$ at the caustic $x = \pm x_s(t)$ for $t > t_*$ is*

$$u_\nu(\pm x_s(t), t) = \pm\frac{x_s(t)}{t} \mp \frac{1}{t}\left(\frac{x_s(t)}{t}\right)^{1/3} + \mathcal{O}(\nu).$$

This matches the solution found from a classical saddle-point analysis obtained by combining (3.2) and (3.7): when $x = x_s(t)$, both saddle points z_0, z_1 coalesce into $z_s = (x_s(t)/t)^{1/3}$. From the asymptotic formula

$$u_\nu(x,t) = \frac{x}{t} - \frac{z_s(x,t)}{t} + \mathcal{O}(\nu)$$

derived in section 3.2, Property 3.5 is verified. Note that this expression is valid only for $t \geq t_* + \varepsilon$, $\varepsilon > 0$.

4. Pole locations.

4.1. Exact pole location at $t = t_*$. From the integral representation (2.2), a Taylor expansion about $x = 0$ can be obtained when $t = t_*$.

Property 4.1. *Let*

$$\mathcal{S}_\nu(z) = \nu^{1/4}\, 2^{-3/4} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(\frac{2n+1}{4})}{\Gamma(2n+1)}\, z^{2n}, \qquad |z| < +\infty,$$

which converges absolutely and uniformly on compact sets for z. Then

$$E_\nu(x, t_*) = \mathcal{S}_\nu\left(\frac{ix}{4t_*}(2\nu)^{-3/4}\right), \qquad |x| < +\infty.$$

Let $x = i\beta$, $\beta \in \mathbb{R}$, $|\beta| < +\infty$; then if we introduce the scaling

$$\beta = \beta(t_*, \nu) = 4t_*(2\nu\mu)^{3/4}, \tag{4.1}$$

we have

$$E_\nu\left(i \cdot 4t_*(2\nu\mu)^{3/4}, t_*\right) = \mathcal{S}_\nu\left(\mu^{3/4}\right). \tag{4.2}$$

Following this scaling, we transform the integral representation of $E_\nu(i\beta, t_*)$ to simplify its analysis. At the inviscid shock time t_*,

$$E_\nu(i\beta, t_*) = \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\nu}\left(\frac{i\beta}{t_*}y - y^4\right)\right\} dy,$$

and the change of variable

$$y \to \left(\frac{\beta}{4t_*}\right)^{1/3} z \tag{4.3}$$

introduces the scaling factor (4.1) between the imaginary part β_n of the zeros a_n and the viscosity ν. This allows us to express $E_\nu(i\beta, t_*)$ in terms of a new function $F(\mu)$, which has the advantage that its saddle points are fixed to the unit disc (thereby making the asymptotic analysis simpler):

$$E_\nu(i\beta, t_*) = \left(\frac{\beta}{4t_*}\right)^{1/3} F\left(\frac{1}{2\nu}\left(\frac{\beta}{4t_*}\right)^{4/3}\right), \quad F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz - z^4)} dz. \tag{4.4}$$

Once the zeros $\{\mu_k\}_{k=1}^{\infty}$ of $F(\mu)$ are found (independent of ν), the poles $\pm a_k(t_*, \nu) = \pm i\beta_k(t_*, \nu)$ of $u_\nu(x, t_*)$ are given by the relation

$$\beta_k(t_*, \nu) = 4t_*(2\nu\mu_k)^{3/4} \quad \forall \nu > 0, \tag{4.5}$$

which was introduced in (4.1). It is important to see that this relation is valid regardless of whether ν is small or β is large. Thus, if we can describe the μ_k's accurately, then the pole locations are known with great precision at t_* (independent of ν). Furthermore, the expansion of $E_\nu(i\beta, t_*)$ as $\nu \to 0^+$ or as $\beta \to +\infty$ is determined by that of $F(\mu)$ as $\mu \to +\infty$. The following theorem is proved in [31].

Theorem 4.2. *The asymptotic expansion of $F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz - z^4)} dz$ as $\mu \to +\infty$ centered about the sector $|\arg \mu| < \pi/2$ is*

$$F(\mu) = \sqrt{\frac{2\pi}{3\mu}} e^{-\frac{3}{2}\mu}\left[\cos\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right) + \mathcal{O}\left(\frac{1}{\mu}\right)\right] \quad \text{as } \mu \to +\infty.$$

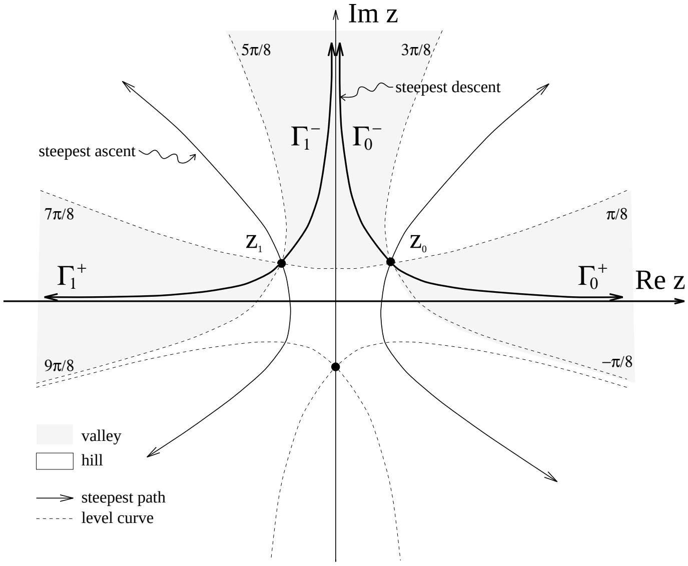


FIG. 4.1. *Hills, valleys, level curves, and steepest paths of the saddle points* $z_0 = e^{i\pi/6}$, $z_1 = e^{i5\pi/6}$ *relevant to the expansion of* $F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz-z^4)}dz$ *as* $\mu \to +\infty$.

Moreover, the kth ordered positive zero μ_k *of* $F(\mu)$ *for* $k \geq 1$ *is given by*

$$\mu_k^{(0)} = \frac{2\pi}{3\sqrt{3}}(k - 1/3), \qquad \mu_k = \mathcal{G}\left(\mu_k^{(0)}\right) + \mathcal{O}\left(\frac{1}{k^6}\right) \quad \textit{as } k \to +\infty,$$
$$\mathcal{G}(\mu) = \mu + \frac{7}{432\mu}\left(1 - \frac{1}{6\mu}\left(1 + \frac{7}{72\mu}\left(1 - \frac{5}{12\mu}\left(1 + \frac{53143}{18900\mu}\right)\right)\right)\right).$$

In addition to this asymptotic description, the first nine values of μ_k are computed numerically in [31] and are listed in section 5. The accuracy of this asymptotic approximation is discussed in [31]; the necessity for such high accuracy will be apparent in section 5. Combining (4.5) and Theorem 4.2, the pole locations at $t = t_*$ are given by the simple form in Property 4.3.

PROPERTY 4.3. *For all* $\nu > 0$,

$$\beta_k(t_*, \nu) = 4t_* \left(\frac{4\nu\pi}{3\sqrt{3}}\right)^{3/4} \cdot \left((k - 1/3)^{3/4} + \mathcal{O}(1/k^{3/4})\right)$$

as $k \to +\infty$.

4.2. Asymptotic analysis of the pole locations $\beta_k(t, \nu)$ as $\nu \to 0^+$ for fixed k and $t \neq t_*$. The saddle-point analysis of $E_\nu(i\beta, t)$ as $\nu \to 0^+$ in the case $t \neq t_*$ is very similar to the one that is described in [31]. There are, again, two equally relevant saddle points which come in a symmetric pair. Note that in this case, their positions are time dependent, and the steepest descent paths are very similar to those displayed in Fig. 4.1, except that the saddle points are either closer together or further apart depending on whether $t < t_*$ or $t > t_*$.

Let $w(z, i\beta) = i\beta z/t + \alpha z^2 - z^4$; $2\alpha = 1/t_* - 1/t < 0$ for $0 < t < t_*$. The saddle points $z_s(\beta;t)$ are the roots of

$$0 = \frac{\partial w}{\partial z}(z, i\beta) = \frac{\beta}{t} i + 2\alpha z - 4z^3. \tag{4.6}$$

Throughout the remainder of the analysis, we write $(\beta;t)$ (as in $z_s(\beta;t)$) to denote that t is to be considered as a parameter. We let $z = -iu$; then u satisfies a cubic equation with real coefficients, namely,

$$u^3 + \frac{\alpha}{2}u - \frac{\beta}{4t} = 0.$$

In order to have some cancellation in the expansion of E_ν to obtain zeros, we need two equally relevant saddle points. Let $x_s(t) = t(2\alpha/3)^{3/2} = i\,(3t_*)^{-3/2}(t_* - t)^{3/2}t^{-1/2}$. Then for $t < t_*$, we find that two of the saddle points come in conjugate pairs only when $|\beta| > |x_s(t)|$, where $\beta = \pm|x_s(t)|$ is the boundary of analyticity of the inviscid solution up to $t = t_*$ (see [32, App. C]). From Cardan's formula for the roots of a cubic polynomial (see [32, App. B]), we find

$$\begin{cases} u_0 = \omega\mathcal{A} + \omega^2\mathcal{B} = -\frac{1}{2}\left(\mathcal{A}+\mathcal{B}\right) + i\frac{\sqrt{3}}{2}\left(\mathcal{A}-\mathcal{B}\right), \\ u_1 = \overline{u_0} = \omega^2\mathcal{A} + \omega\mathcal{B} = -\frac{1}{2}\left(\mathcal{A}+\mathcal{B}\right) - i\frac{\sqrt{3}}{2}\left(\mathcal{A}-\mathcal{B}\right), \\ u_2 = \mathcal{A}+\mathcal{B}, \end{cases} \tag{4.7}$$

where $\omega = e^{2\pi i/3}$. Since $x_s^2 = -|x_s^2| < 0$ for $t < t_*$, we find

$$\begin{aligned} &\begin{cases} \mathcal{A}(\beta;t) = (8t)^{-1/3}\ \sqrt[3]{\beta + \sqrt{\beta^2 + x_s^2}} > 0 \\ \mathcal{B}(\beta;t) = (8t)^{-1/3}\ \sqrt[3]{\beta - \sqrt{\beta^2 + x_s^2}} > 0 \end{cases} && \text{for } \boldsymbol{\beta > |x_s(t)|}, \\ &\begin{cases} \mathcal{A}(\beta;t) = -\,(8t)^{-1/3}\ \sqrt[3]{-\beta + \sqrt{\beta^2 + x_s^2}} < 0 \\ \mathcal{B}(\beta;t) = -\,(8t)^{-1/3}\ \sqrt[3]{-\beta - \sqrt{\beta^2 + x_s^2}} < 0 \end{cases} && \text{for } \boldsymbol{\beta < -|x_s(t)|}, \end{aligned} \tag{4.8}$$

so

$$z_s(-\beta;t) = -z_s(\beta;t). \tag{4.9}$$

Here we are taking the real positive branches of the square roots and cube roots in $\mathcal{A}$ and $\mathcal{B}$. We have also used the relation

$$|x_s\,(t)\,| = t\left(\frac{2|\alpha|}{3}\right)^{3/2} = t\left(-\frac{2\alpha}{3}\right)^{3/2} > 0, \tag{4.10}$$

where we are taking the positive branch of $z^{3/2}$ for $z > 0$. Moreover, when choosing the branches of the rational functions $\mathcal{A}$ and $\mathcal{B}$, one must make sure that they satisfy the relation $\mathcal{A}\cdot\mathcal{B} = -\alpha/6 > 0$ (see [32, App. B]). In terms of the original variable $z = -iu$, after separation of the real and imaginary parts, we have

$$\begin{cases} z_0 = \frac{\sqrt{3}}{2}(\mathcal{A}-\mathcal{B}) + \frac{i}{2}(\mathcal{A}+\mathcal{B}), \\ z_1 = -\overline{z_0} = \frac{\sqrt{3}}{2}(\mathcal{B}-\mathcal{A}) + \frac{i}{2}(\mathcal{A}+\mathcal{B}), \\ z_2 = -i(\mathcal{A}+\mathcal{B}). \end{cases} \tag{4.11}$$

Since we are only looking at values of $|\beta| > |x_s|$, the steepest paths and level curves look almost like the case $t = t_*$ described in Fig. 4.1 except that the saddle points have moved closer together, yet preserve the symmetry of Fig. 4.1. The path deformation is justified in the same way (see [31] for more details). The saddle points come in symmetric pairs that satisfy $z_0 = -\overline{z_1}$ for all $t > 0$. We have

$$0 = \frac{\partial w}{\partial z}\big(z_s(\beta;t), i\beta\big) = \frac{\beta}{t} i + 2\alpha z_s - 4z_s^3, \tag{4.12a}$$

$$0 = \frac{1}{4} z_s \frac{\partial w}{\partial z}\big(z_s(\beta;t), i\beta\big) = \frac{i\beta}{4t} z_s + \frac{\alpha}{2} z_s^2 - z_s^4, \tag{4.12b}$$

$$w\big(z_s(\beta;t), i\beta\big) = \frac{i\beta}{t} z_s + \alpha z_s^2 - z_s^4. \tag{4.12c}$$

Equation (4.12a) gives (4.12b), which combined with (4.12c) gives

$$w\big(z_s(\beta;t), i\beta\big) = \frac{3i\beta}{4t} z_s + \frac{\alpha}{2} z_s^2. \tag{4.13}$$

Since for $s = 0, 1$

$$z_s(\beta;t) = (-1)^s \frac{\sqrt{3}}{2}(\mathcal{A} - \mathcal{B}) + \frac{i}{2}(\mathcal{A} + \mathcal{B}), \tag{4.14a}$$

$$w_{zz}(z_s, i\beta) = 2\alpha - 12z_s^2, \quad w(z_0, i\beta) = \overline{w(z_1, i\beta)}, \tag{4.14b}$$

so

$$\Re\, w\big(z_s(\beta;t), i\beta\big) = \frac{\alpha}{4}(\mathcal{A}^2 + \mathcal{B}^2) - \frac{3\beta}{8t}(\mathcal{A} + \mathcal{B}) - \frac{\alpha^2}{6}, \tag{4.15a}$$

$$\Im\, w\big(z_s(\beta;t), i\beta\big) = (-1)^s \frac{\sqrt{3}}{8} \cdot (\mathcal{A} - \mathcal{B}) \cdot (3\beta/t + 2\alpha(\mathcal{A} + \mathcal{B}))\,, \tag{4.15b}$$

$$\theta(z_s(\beta;t), t) = \arg(-w_{zz}(z_s, i\beta)) = (-1)^s \arg(6z_0^2 - \alpha). \tag{4.15c}$$

Using a standard steepest descents analysis (see [31, 39] for example), we find that

$$E_\nu(i\beta, t) = \sum_{s=0,1} \sqrt{\frac{-4\pi\nu}{w_{zz}(z_s, i\beta)}} e^{w(z_s, i\beta)/2\nu} \left(1 + \mathcal{O}(\nu)\right) \quad \text{as } \nu \to 0^+.$$

We can further simplify the expansion using the actual value of $\sqrt{6z_s^2 - \alpha}$. Indeed, since $z_s(\beta;t) = (-1)^s \frac{\sqrt{3}}{2}(\mathcal{A} - \mathcal{B}) + \frac{i}{2}(\mathcal{A} + \mathcal{B}) = e^{i\pi/6}\mathcal{A} + e^{i5\pi/6}\mathcal{B}$ and $\mathcal{A} \cdot \mathcal{B} = -\alpha/6$,

$$6z_s^2 - \alpha = 3\sqrt{(\mathcal{A}^2 + \mathcal{B}^2 - \alpha)^2 + 3(\mathcal{A}^2 - \mathcal{B}^2)}\, e^{i\theta(z_s(\beta;t),t)}, \tag{4.16a}$$

$$\theta(z_s(\beta;t), t) = \arg(6z_s^2 - \alpha) = (-1)^s \tan^{-1}\left(\sqrt{3} \cdot \frac{\mathcal{A}^2 - \mathcal{B}^2}{\mathcal{A}^2 + \mathcal{B}^2 + \alpha/3}\right), \tag{4.16b}$$

where in (4.16b) we are taking the branch of $\tan^{-1} x$ for which $|\tan^{-1} x| < \pi/2$. Reproducing a similar analysis for $t > t_*$, we have the following result.

THEOREM 4.4. *The asymptotic expansion of $E_\nu(i\beta, t)$ as $\nu \to 0^+$ is*

$$E_\nu(i\beta, t) = \sqrt{\frac{2\pi\nu}{|6z_s^2 - \alpha|}} \exp\left\{\frac{1}{2\nu} \Re\, w\big(z_s(\beta;t), i\beta\big)\right\} \times \left[\cos\Big(\frac{1}{2\nu} \Im\, w\big(z_0(\beta;t), i\beta\big) - \frac{1}{2}\theta(z_0(\beta;t), t)\Big) + \mathcal{O}\left(\nu\right)\right],$$

where

$$z_s(\beta;t) = (-1)^s \frac{\sqrt{3}}{2}(\mathcal{A}-\mathcal{B}) + \frac{i}{2}(\mathcal{A}+\mathcal{B}) \qquad \textit{for } s = 0, 1,$$

and $\mathcal{A}$ *and* $\mathcal{B}$ *are given by*

$$\mathcal{A}(\beta;t) = (8t)^{-1/3}\ \sqrt[3]{\beta + \sqrt{\beta^2 + x_s^2}}$$
$$\mathcal{B}(\beta;t) = \begin{cases} 0 < t < t_* & (8t)^{-1/3}\ \sqrt[3]{\beta - \sqrt{\beta^2 + x_s^2}} > 0 \qquad \beta > |x_s|, \\ t > t_* & -(8t)^{-1/3}\ \sqrt[3]{-\beta + \sqrt{\beta^2 + x_s^2}} < 0 \qquad \beta > 0. \end{cases}$$

For $\beta < 0$, $z_s(\beta;t)$ is defined by the odd parity condition $z_s(\beta;t) = -z_s(-\beta;t)$. Letting $t \to t_*$ in Theorem 4.4, we obtain $\theta(z_s(\beta;t_*), t_*) = (-1)^s \pi/3$, and

$$\Re\, w\big(z_s(\beta;t_*), i\beta\big) = -\frac{3}{2}\left(\frac{\beta}{4t_*}\right)^{4/3},$$
$$\Im\, w\big(z_s(\beta;t_*), i\beta\big) = (-1)^s \frac{3\sqrt{3}}{2}\left(\frac{\beta}{4t_*}\right)^{4/3}.$$

For small ν, the poles β_k are approximated by the roots of the equation

$$\frac{1}{2\nu}\Im\, w\big(z_0(\beta;t), i\beta\big) - \frac{1}{2}\theta(z_0(\beta;t), t) = \left(k - \frac{1}{2}\right)\pi, \quad k \in \mathbb{N}^*, \tag{4.17}$$

with the convention that $\beta_{-k} \equiv -\beta_k$. Since $|\theta(z_0(\beta_k;t),t)| < \pi$ for all $\beta_k \in \mathbb{R}$, the limiting behavior of the poles is given by $\Im\, w\big(z_0(\beta;t), i\beta\big) = 0$. Recall from (4.15b) that

$$\Im w_s = \Im\, w\big(z_s(\beta;t), i\beta\big) = (-1)^s \frac{\sqrt{3}}{8} \cdot (\mathcal{A}-\mathcal{B}) \cdot (3\beta/t + 2\alpha(\mathcal{A}+\mathcal{B})),$$

so

$$\Im w_s = 0 \Leftrightarrow \begin{cases} \text{either } \mathcal{A} = \mathcal{B} \text{ or} \\ 3\beta/t + 2\alpha(\mathcal{A}+\mathcal{B}) = 0. \end{cases} \tag{4.18}$$

For $0 < t \le t_*$, $\alpha \le 0$: if $\beta \ge |x_s|$ then $\mathcal{A} > 0$, $\mathcal{B} > 0$; if $\beta \le |x_s|$ then $\mathcal{A} < 0$, $\mathcal{B} < 0$. We rewrite the second equation as $(\mathcal{A}+\mathcal{B})^3 = -\left(3\beta/2\alpha t\right)^3$, which, after expanding the left-hand side and using the fact that $\mathcal{A}\cdot\mathcal{B} = -\alpha/6$, reduces to $\beta = \pm|x_s(t)|$. The same conclusion is reached from the first equation $\mathcal{A} = \mathcal{B}$. Thus, for $0 < t \le t_*$, $\Im w_s = 0$ only if $\beta = \pm|x_s|$. Let $\hat{\beta} \ll \beta$. Then, re-substituting $\beta = |x_s(t)| + \hat{\beta}$ into the expansion for $E_\nu(i\beta, t)$ in Theorem 4.4 and reproducing an analysis which is similar to the one described in [31] (i.e., inverting the asymptotic series expansion), we find that the error term is $\mathcal{O}((k\nu)^{3/4})$ as $\nu \to 0^+$ for fixed k. Thus, we can write that $\beta_{\pm k}(t,\nu) = \pm|x_s| + \mathcal{O}((k\nu)^{3/4})$ as $\nu \to 0^+$ for fixed k. Similarly, for $t \ge t_*$, $\alpha \ge 0$, $\beta > 0 \Rightarrow \mathcal{A}-\mathcal{B} > 0$ and $\beta < 0 \Rightarrow \mathcal{A}-\mathcal{B} < 0$; hence, $\Im w_s = 0 \Leftrightarrow \beta = 0$ as a result of setting $3\beta/t + 2\alpha(\mathcal{A}+\mathcal{B}) = 0$. Since $\Im x_s(t) = 0$ for $t \ge t_*$, we have proved the following (see Fig. 5.12).

Corollary 4.5. *For all* $t > 0$ *and fixed* k, *the asymptotic behavior of the poles* $x = \pm a_k(t,\nu) = \pm i\beta_k(t,\nu)$ *is given by*

$$\beta_k(t,\nu) = \Im x_s(t) + \mathcal{O}((k\nu)^{3/4}) \qquad \textit{as } \nu \to 0^+.$$

Since $\Im x_s(t) = 0$ *for* $t \geq t_*$,

$$\beta_k(t,\nu) = \mathcal{O}((k\nu)^{3/4}) \qquad \textit{as } \nu \to 0^+$$

for fixed k. *Of particular interest is the modulus of the first ordered pole* $\beta_1(t,\nu)$ *which governs the time evolution of the width of the analyticity strip of the viscous solution.*

Following the exact same steps in the proof of Corollary 4.5, one can show the following corollary.

COROLLARY 4.6. *For all* $t > 0$ *and fixed* ν, *the asymptotic behavior of the poles* $x = \pm a_k(t,\nu) = \pm i\beta_k(t,\nu)$ *is*

$$\beta_k(t,\nu) = \mathcal{O}((k\nu)^{3/4}) \qquad \textit{as } k \to +\infty.$$

5. Numerics.

5.1. Finite difference approximation, asymptotic approximation, and pole expansion. We present a numerical scheme which enables us to solve (1.1) for moderately small values of ν. The procedure is sometimes referred to as the method of lines and consists in using a centered difference operator in space while time-marching with a Runge–Kutta scheme. The method is implemented on the interval $I = [0, 1/2]$ with boundary conditions

$$u_\nu(0,t) = u_\nu(1/2,t) = 0. \tag{5.1}$$

The boundary condition $u_\nu(1/2,t) = 0$ is chosen to be consistent with the value of the inviscid solution $u(1/2,t) = 0$. Thus, we can expect the difference approximation to be consistent with the initial (boundary) value problem for small ν. Two different initial conditions are also used:

$$u(x,0) = u_\nu(x,0) = 4x^3 - \frac{x}{t_*}, \tag{5.2a}$$

$$u_\nu(x,t_*) = \frac{x}{t_*} - \sum_{n=1}^{\infty} \frac{4\nu x}{x^2 + \beta_n^2(t_*,\nu)}. \tag{5.2b}$$

Throughout the numerics we use the parameter value $t_* = 1$. If the second condition is used, then the pole positions at $t = t_*$ are specified by the asymptotic estimate presented in Theorem 4.2. This estimate is used for all values of μ_n for $10 \leq n \leq N$:

$$\begin{cases} \beta_n(t_*,\nu) = 4t_*(2\nu\mu_n)^{3/4}, \\ \mu_n = G(\mu_n^{(0)}), \quad \mu_n^{(0)} = \frac{2\pi}{3\sqrt{3}}(n - 1/3), \quad n \geq 10, \\ G(\mu) = \mu + \frac{7}{432\mu}\left(1 - \frac{1}{6\mu}\left(1 + \frac{7}{72\mu}\left(1 - \frac{5}{12\mu}\left(1 + \frac{53143}{18900\mu}\right)\right)\right)\right). \end{cases} \tag{5.3}$$

For $1 \leq n \leq 9$, we use the numerical values found in [31, Table 3], under the column "Numerical roots":

$$\mu_1 = 0.8221037147, \quad \mu_2 = 2.0226889660, \quad \mu_3 = 3.2292915284, \tag{5.4}$$

$$\mu_4 = 4.4372464748, \quad \mu_5 = 5.6457167459, \quad \mu_6 = 6.8544374340, \tag{5.5}$$

$$\mu_7 = 8.0632985369, \quad \mu_8 = 9.2722462225, \quad \mu_9 = 10.4812510479. \tag{5.6}$$

Let

$$\begin{aligned} &u_j = u_\nu(j * \Delta x, t), \quad Ev_j = v_{j+1}, \quad E^p v_j = v_{j+p}, \\ &D_+ = (E - E^0)/\Delta x, \quad D_- = (E^0 - E^{-1})/\Delta x, \quad D_0 = (D_+ + D_-)/2. \end{aligned}$$

One then solves the system of $N_x - 1$ equations using a Runge–Kutta 4–5 scheme (which we refer to as RK45):

$$\begin{cases} du_j/dt = -D_0(u_j^2/2) + \nu D_+ D_- u_j, \quad j = 1, \dots, N_x - 1, \\ u_{j=0} = u_\nu(0,t) = 0, \quad u_{j=N_x} = u_\nu(1/2,t) = 0, \end{cases} \tag{5.7}$$

where N_x is the number of gridpoints (and gridfunctions) and $N_x * \Delta x = 1/2$. Typically, the mesh size we use is $\Delta x = .25 \times 10^{-2}$ and $N_x = 200$ gridpoints. The time stepping restrictions depend on the size of ν and on how far in time one wants to go. For example, if the final time is $t = t_* = 1$, whether $\nu = 10^{-2}$ or $\nu = 10^{-3}$ it suffices to use $\Delta t = .25 \times 10^{-2}$, $N_t = 400$ RK45 steps. However, for $\nu = 10^{-2}$, if one wants to go as far as $t = 2$, for reasons of stability one needs to use a smaller time step such as $\Delta t = 10^{-3}$, $N_t = 2,000$. The domain of integration is $(x,t) \in [0, 1/2] \times [0, T]$, where $T = 1$ or $T = 2$. Then, due to the odd parity of the solution, we reflect symmetrically for $x \in [-1/2, 0]$ according to the rule $u_\nu(-x,t) = -u_\nu(x,t)$. This finite difference scheme is used in order to compare the predictions obtained from the pole expansion and the pole dynamics in section 5.2.

5.2. Numerical pole dynamics.

We investigate the motion of the simple poles of $u_\nu(x,t)$ by solving the truncated CDS and by starting with initial data for the poles at $t = t_*$. The poles of $u_\nu(x,t)$ are located at $\pm a_n(t,\nu) = \pm i\sqrt{\nu\gamma_n(t,\nu)}$, where the variables $\gamma_n(t,\nu) > 0$ satisfy the system (cf. Property 2.2)

$$\begin{cases} \dfrac{\dot{\gamma}_n}{2} = \dfrac{\gamma_n}{t} + 1 - 4\,\gamma_n \sum_{l \neq n} \dfrac{1}{\gamma_l - \gamma_n} \\ \gamma_n(t_*,\nu) = (4t_*)^2\,(2\mu_n)^{3/2}\,\sqrt{\nu} \end{cases} \quad \forall n \in \mathbb{N}. \tag{5.8}$$

In order to solve this system we use the asymptotic estimate for μ_n presented in (5.3) and the numerical values of (5.5). We are mainly interested in describing the motion of the first pole $a_1(t,\nu) = i\beta_1(t,\nu) \in i\mathbb{R}$; this amounts to describing the time evolution of the width of the strip of analyticity of the solution $u_\nu(t,x)$. The imaginary part of the poles $\beta_n(t,\nu)$ is recovered using the relation $\beta_n(t,\nu) = \sqrt{\nu\gamma_n(t,\nu)}$. We plot the evolution of $\beta_n(t,\nu), n = 1, \dots, 4$ for different values of ν. We use N poles in the computations, i.e., β_1 through β_N where $N \times 10^{-4}$ varies from $.1, .5, 1, 2.5, 5$. That is, we consider the truncated system

$$\begin{cases} \dfrac{\dot{\gamma}_n}{2} = \dfrac{\gamma_n}{t} + 1 - 4\gamma_n \sum_{\substack{l=1 \\ l \neq n}}^{N} \dfrac{1}{\gamma_l - \gamma_n} \\ \gamma_n(t_*,\nu) = (4t_*)^2(2\mu_n)^{3/2}\sqrt{\nu} \end{cases} \quad \forall n = 1, \dots, N.$$

In order to accelerate the computation of the slowly converging pole expansions which require $\mathcal{O}(N^2)$ operations

$$\sum_{\substack{l=1 \\ l \neq n}}^{N} \frac{1}{\gamma_l - \gamma_n} \qquad \forall n = 1, \dots, N, \tag{5.9}$$

we use a multipole algorithm developed and implemented by Greengard and Rokhlin [14, 21] which reduces the computational complexity to $\mathcal{O}(N \log N)$. A fourth/fifth-order Runge–Kutta–Fehlberg scheme with automatic step-size control is used (the same one that is used for the finite difference scheme/method of lines computations of the previous section). Since the initial data is specified at $t = t_* = 1$, we can solve the system forward and backwards in time starting from $t = 1$. The typical bound on the relative error in the computation is $10^{-8} < |(x_4 - x_5)/x_5| < 10^{-4}$, where x_4 and x_5 are, respectively, the 4th- and 5th-order estimates of $\gamma_1(t,\nu)$. Once this error criterion is met, we recover the pole location via the relation $a_n(t,\nu) = i\sqrt{\nu\gamma_n(t,\nu)}$. The justification of the numerics is the most difficult aspect of this simulation because one must justify the convergence of the method as both the number of poles increases and the time step is refined. The time-step control is automatically determined by the local relative tolerance (LRT $= |(x_4 - x_5)/x_5|$) test on the 4th- and 5th-order approximations of the first ordered pole (the one closest to the origin). Thus, one cannot fix the time stepping; rather, one can have a subtle control on it by reducing this tolerance. Typically, we fix the number of poles to $N = 50,000$ and vary the tolerance on the successive intervals $10^{-10} < LRT < 10^{-6}$, $10^{-8} < LRT < 10^{-4}$, and $10^{-6} < LRT < 10^{-2}$. Then we fix the tolerance at the highest reasonable level $10^{-8} < LRT < 10^{-4}$ and vary the number of poles where $N \times 10^{-4}$ is either $.1, .5, 1, 2.5$, or 5. We see that the time step barely affects the convergence of the method. Thus, the main difficulty in this procedure arises from the slow convergence of the pole interaction (5.9) that is present in the CDS.

5.2.1. Exact solution for the two-pair pole-dynamics test. In order to verify the accuracy of the numerical pole dynamics, we implement the numerical method described in the previous section for the case where there are only four poles (two pairs). In this case, one can explicitly solve the resulting system as follows: let $a_n = i\beta_n$. That is, replace β_n by $-ia_n$ in Property 2.2 so that the two pairs of poles $\{(-a_1, a_1), (-a_2, a_2)\}$ and $\{(-\kappa_1, \kappa_1), (-\kappa_2, \kappa_2)\}$ satisfy, under the transformation $\kappa_n = a_n^2/\nu$, the equivalent systems

$$\left\{\begin{array}{l} \dot{a}_1 = a_1/t - \nu/a_1 - 4\nu a_1/(a_1^2 - a_2^2), \\ \dot{a}_2 = a_2/t - \nu/a_2 + 4\nu a_2/(a_1^2 - a_2^2), \end{array}\right. \Longleftrightarrow \left\{\begin{array}{l} \dot{\kappa}_1/2 = \kappa_1/t - 1 - 4\kappa_1/(\kappa_1 - \kappa_2), \\ \dot{\kappa}_2/2 = \kappa_2/t - 1 + 4\kappa_2/(\kappa_1 - \kappa_2). \end{array}\right.$$

Introduce a set of new variables $\{\Theta_1,\Theta_2\}$ defined by

$$\left\{\begin{array}{l} \Theta_1 = \kappa_1 + \kappa_2, \\ \Theta_2 = \kappa_1 - \kappa_2, \end{array}\right. \Longleftrightarrow \left\{\begin{array}{l} \kappa_1 = (\Theta_1 + \Theta_2)/2, \\ \kappa_1 = (\Theta_1 - \Theta_2)/2. \end{array}\right.$$

Then it is easy to show that $\{\Theta_1,\Theta_2\}$ satisfy the coupled system of nonlinear ODEs

$$\left\{\begin{array}{l} \dot{\Theta}_1 - 2\Theta_1/t = -12, \\ \dot{\Theta}_2 - 2\Theta_2/t = -8\Theta_1/\Theta_2. \end{array}\right. \tag{5.10}$$

We further introduce a new variable denoted by $\phi_2 = \Theta_2^2$ which in turn satisfies the linear ODE

$$\dot{\phi}_2 - 4\phi_2/t = -16\Theta_1. \tag{5.11}$$

We use as initial data the position of the poles $a_1(t_*,\nu) = i\beta_1(t_*,\nu)$ and $a_2(t_*,\nu) = i\beta_1(t_*,\nu)$, where $\beta_1(t_*,\nu)$ and $\beta_2(t_*,\nu)$ are given in (5.3) and (5.4). Thus, we have

$$\left\{\begin{array}{l} \Theta_1^* = \Theta_1(t_*,\nu) = \kappa_1(t_*,\nu) + \kappa_2(t_*,\nu) = a_1^2(t_*,\nu)/\nu + a_2^2(t_*,\nu)/\nu, \\ \phi_2^* = \phi_2(t_*,\nu) = \Theta_2^2(t_*,\nu) = (a_1^2(t_*,\nu)/\nu - a_2^2(t_*,\nu)/\nu)^2. \end{array}\right. \tag{5.12}$$

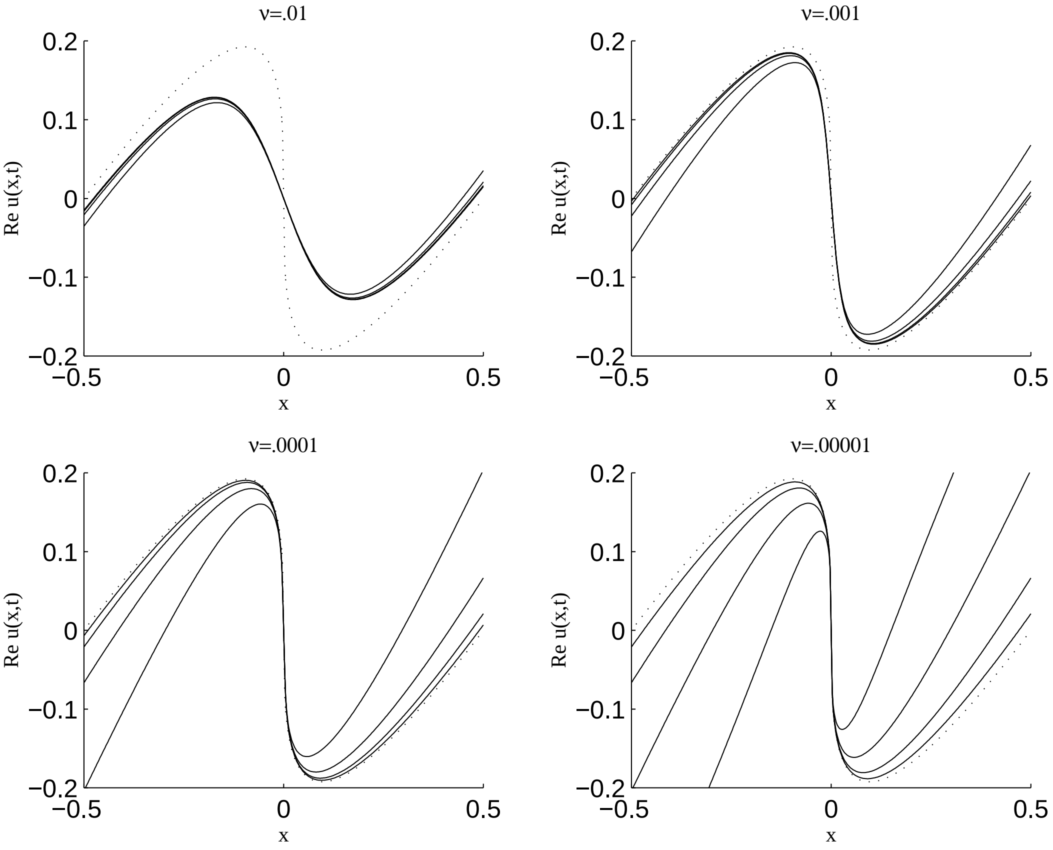


FIG. 5.1. *Convergence of the pole expansion as $N \rightarrow +\infty$ of the solution $u_\nu(x,t_*) = x/t_* - \Sigma_{n=1}^{N} 4\nu x/(x^2+\beta_n^2(t_*,\nu))$ with varying number of poles ranging from $N = 10^3, 10^4, 10^5, 10^6$ poles for $\nu = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$. The dotted curve is computed from the inviscid solution at $t = t_* = 1$ by $u(x,t_*) = x/t_* - (x/4t_*)^{1/3}$. For $\nu = 10^{-2}$, the inviscid solution and the pole expansion $u_\nu(x,t_*)$ do not agree because the viscosity is large enough that the solution has started decaying earlier (see comments on the turn-around time of the poles and their relation to the decay of the solution).*

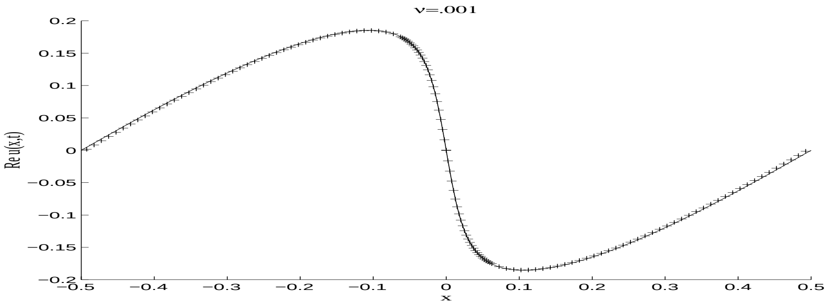


FIG. 5.2. *Comparison of the solution reconstruction at $t = t_* = 1$ from the pole expansion $u_\nu(x,t_*) = x/t_* - \Sigma_{n=1}^{N} 4\nu x/(x^2+\beta_n^2(t_*,\nu))$ with $N = 10^6$ poles and the finite difference scheme (method of lines) for $\nu = 10^{-3}$. Mesh size: $N_x = 200$ points, $\Delta x = .25 * 10^{-2}$, $N_t = 400$ RK45 time steps with $\Delta t = .25 * 10^{-2}$. Pole expansion (+) at $t = 1$ overlaps finite difference approximation in solid curve.*

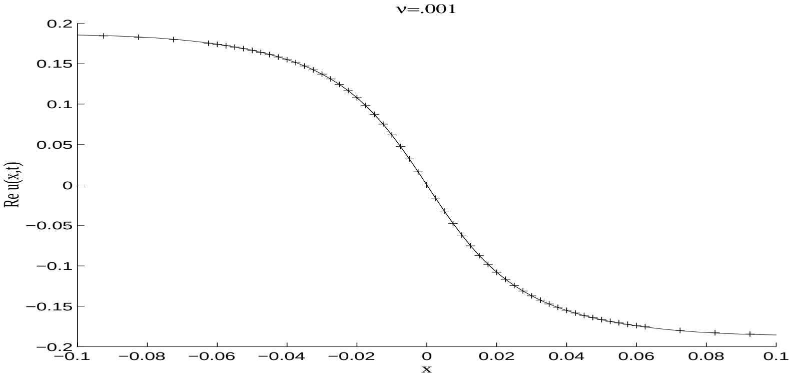


FIG. 5.3. *Closeup of Fig.* 5.2 *in* $[-.1, .1]$.

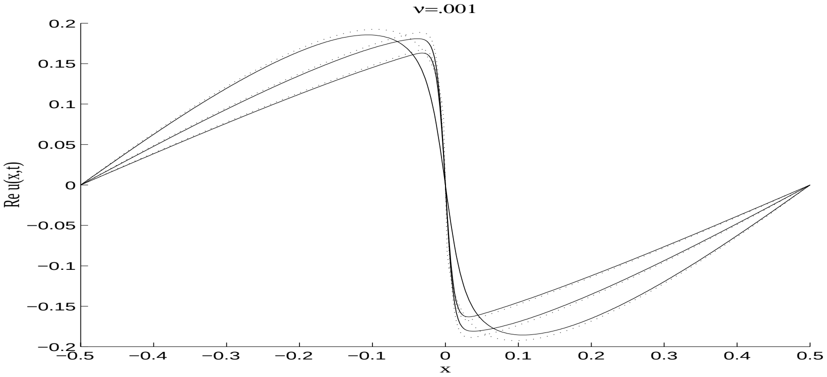


FIG. 5.4. *Comparison of finite difference scheme and saddle-point method for* $\nu = 10^{-3}$ *at* $t = 1, 1.5, 2$. *Solid curves: finite difference scheme with* $N_x = 200$ *points,* $\Delta x = .25 * 10^{-2}$, $N_t = 800$ $RK45$ *time steps with* $\Delta t = .25 * 10^{-2}$. *Dotted curves: saddle-point approximation overshooting the finite difference approximation.*

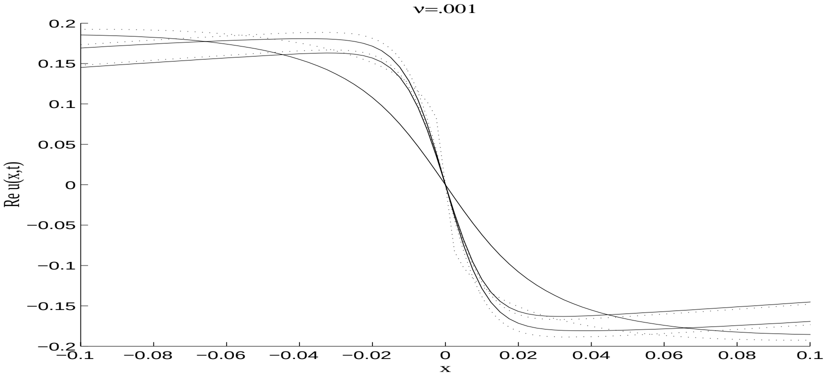


FIG. 5.5. *Closeup of Fig.* 5.4 *in* $[-.1, .1]$.

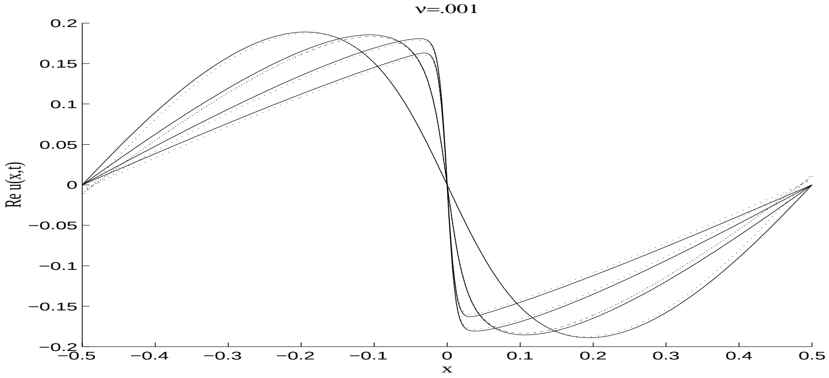


FIG. 5.6. *Comparison of the finite difference approximation (solid) and the pole dynamics (dotted) for* $\nu = 10^{-3}$ *at* $t = .5, 1, 1.5, 2$*. Finite difference mesh size:* $N_x = 200$ *points,* $\Delta x = .25 * 10^{-2}$*,* $N_t = 2,000$ *RK45 steps with* $\Delta t = .5 * 10^{-2}$*. Pole dynamics:* $N = 5 * 10^4$ *poles,* $10^{-8} < LRT < 10^{-4}$*, typical time step* $\Delta t = .05$*,* $N_t = 45$ *RK45 time steps (*25 *steps backward and* 20 *steps forward from* $t = t_*$*).*

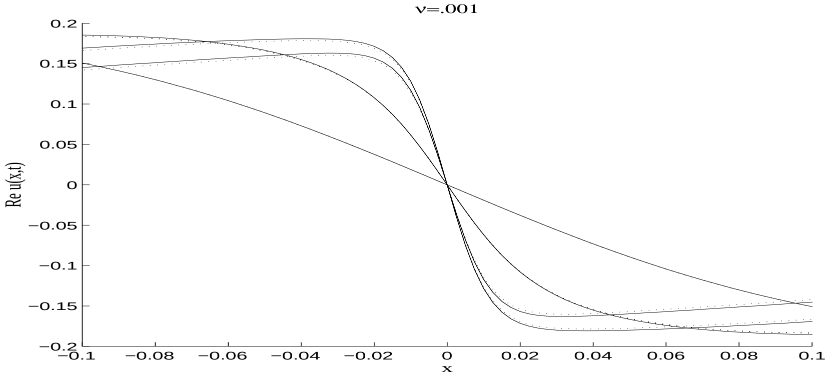


FIG. 5.7. *Closeup of Fig.* 5.6 *in* $[-.1, .1]$.

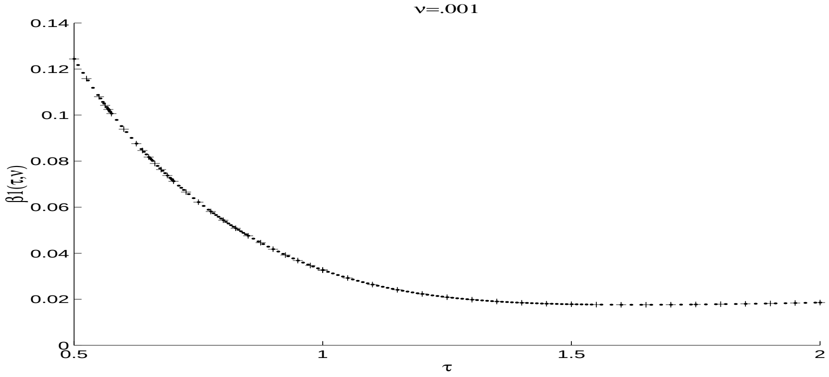


FIG. 5.8. $\beta_1(t,\nu)$ *vs. t. Time evolution in* $\mathbb{R}$ *of the width of the analyticity strip* $\beta_1(t,\nu)$ *for* $\nu = 10^{-3}$ *and* $N = 5 \times 10^4$ *poles.* $t_{initial} = t_* = 1$ *and* $t \in [.5, 2]$*. (+):* $\Delta t = .05$*; dots (.):* $\Delta t = .01$*. Both curves are indistinguishable.*

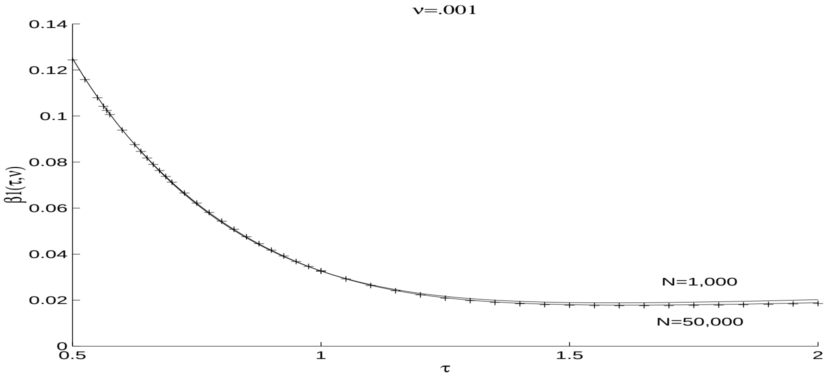


FIG. 5.9. *Closeup of Fig.* 5.8 *for* $t \in [1, 2]$.

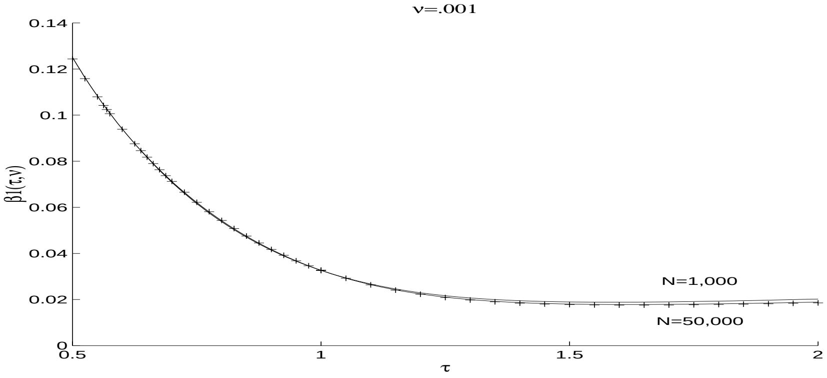


FIG. 5.10. $\beta_1(t, \nu)$ *vs.* t. *Comparison of pole number simulations for* $\nu = 10^{-3}$ *and* $N = .1, .5, 1, 2.5, 5 \times 10^4$ *poles. (+):* $N = 5 \times 10^4$ *poles; (solid):* $N = .1, .5, 1, 2.5 \times 10^4$ *poles. Differences appear more clearly in the closeup in Fig.* 5.11.

Solving the IVP consisting of the first equation in system (5.10) and equations (5.11) and (5.12), we find that

$$\left\{ \begin{array}{l} \Theta_1(t,\nu) = (t/t_*)^2 \Theta_1^* - 12t(t-t_*)/t_*, \\ \phi_2(t,\nu) = (t/t_*)^4 \left(\phi_2^* - 16t_*(t-t_*)(t\Theta_1^* - 6tt_* + 6t_*^2)/t^2\right). \end{array} \right. \tag{5.13}$$

Taking $t_* = 1$, $\nu = .001$, we use (5.13), a straightforward numerical integration scheme using RK45 and RK45 together with the multipole algorithm in which we set to zero all coefficients pertaining to $a_n, n \geq 3$. We find common values for all three methods at $t = 1.25$:

$$\left\{ \begin{array}{l} a_1(t = 1.25, \nu = .001) = 0.0408023705 * i, \\ a_2(t = 1.25, \nu = .001) = 0.1009178717 * i. \end{array} \right. \tag{5.14}$$

Computing the differences between the exact values of a_1 and a_2 and the predictions obtained from the Runge–Kutta schemes (with and without the multipole algorithm), we find that these predictions are of the order of $\mathcal{O}(10^{-10})$, which is consistent with the expected 4th-order accuracy of such numerical schemes.

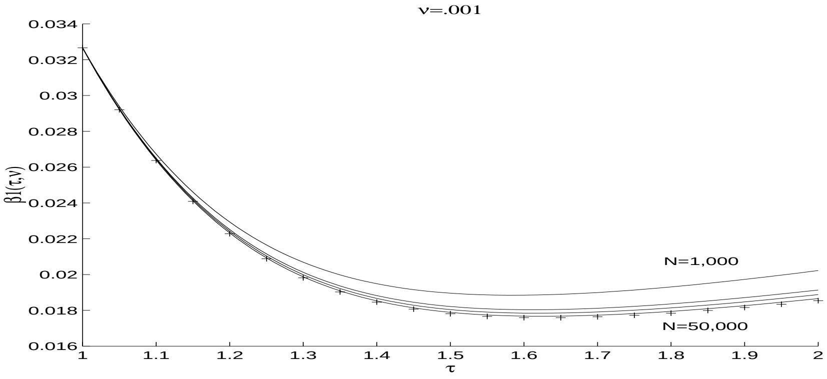


FIG. 5.11. *Closeup of Fig.* 5.10 *for* $t \in [1, 2]$. *Turn-around time at* $t_u \approx 1.62$.

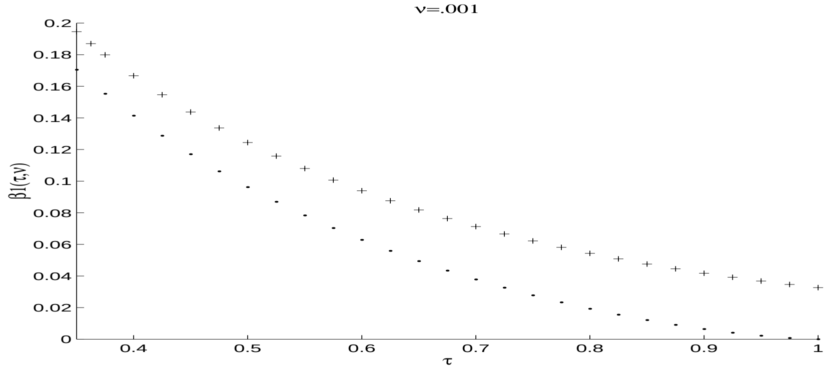


FIG. 5.12. *Comparison of the time evolutions of* $\beta_1(t, \nu = 10^{-3})$ (+) *and* $x_s(t)$ (.) *for* $t \in [.35, t_* = 1]$. *The pole dynamics are the same as Fig.* 5.8 *with* $N = 5 \times 10^4$ *poles. This illustrates the asymptotic relation* $\beta_1(t, \nu) = \Im x_s(t) + \mathcal{O}(\nu^{3/4})$ *as* $\nu \to 0^+$ *when* $t \leq t_*$ *(see Corollary* 4.5*).*

5.3. Figures, descriptions, and comparisons. In Fig. 5.1, we illustrate the "slow" convergence of the pole expansion as the viscosity decreases. In particular, for $\nu = 10^{-4}$ and 10^{-5}, we can compare the inviscid solution given by (see [32, App. C])

$$u(x, t_*) = \frac{x}{t_*} - \left(\frac{x}{4t_*}\right)^{1/3} \tag{5.15}$$

to the pole expansion and expect good agreement between the two. For ν very small, we see that even for a very large number of poles ($N = 10^6$) the tails of the pole expansion still do not match the true solution, which is expected to be very close to the inviscid one. In each of these figures there are five curves, four of which are computed from the pole expansion for an increasing number of poles $N = 10^3, 10^4, 10^5, 10^6$. The fifth (dotted curve) is the inviscid solution at t_*.

In Figs. 5.2 and 5.3 we present comparisons between the finite difference scheme and the pole expansion ($N = 10^6$ poles) at the fixed time t_*. For the finite difference scheme, we use $N_x = 200$ points, $\Delta x = .25 * 10^{-2}$, $N_t = 400$ RK45 steps with $\Delta t = .25 * 10^{-2}$.

In Figs. 5.4 and 5.5 we present comparisons between the finite difference scheme

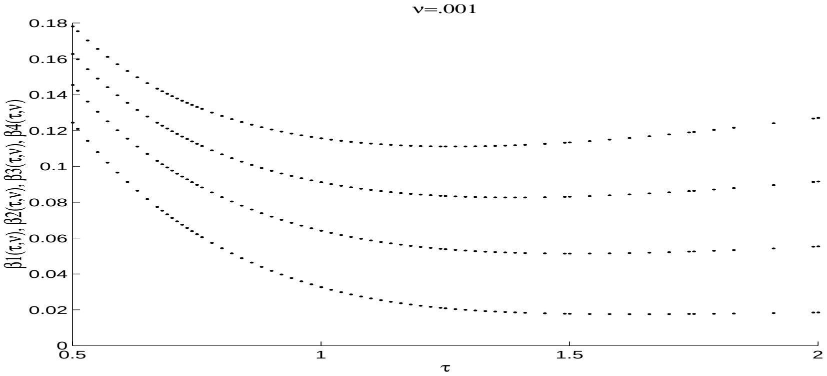


FIG. 5.13. $\beta_j(t,\nu)$ *vs.* t *for* $j = 1,\dots,4$. $\nu = 10^{-3}$ *and* $N = 5\times 10^4$ *poles. Same parameters as in Fig.* 5.8. *Turn-around times at* $t_u \approx 1.62$, $t_u \approx 1.51$, $t_u \approx 1.39$, $t_u \approx 1.27$ *for* $\beta_j(t,\nu), j = 1,\dots,4$.

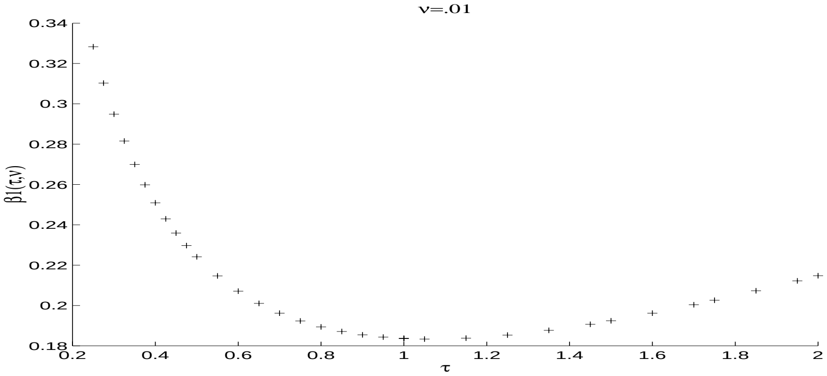


FIG. 5.14. $\beta_1(t,\nu)$ *vs.* t. *Time evolution in* $\mathbb{R}$ *of the width of the analyticity strip* $\beta_1(t,\nu)$ *for* $\nu = 10^{-2}$ *and* $N = 5\times 10^4$ *poles.* $t_{initial} = t_* = 1$ *and* $t \in [.25, 2]$. $N_{steps} = 32$ *(*20 *steps backward and* 12 *steps forward from* $t = t_*$*). Local relative tolerance:* $10^{-10} < LRT < 10^{-6}$.

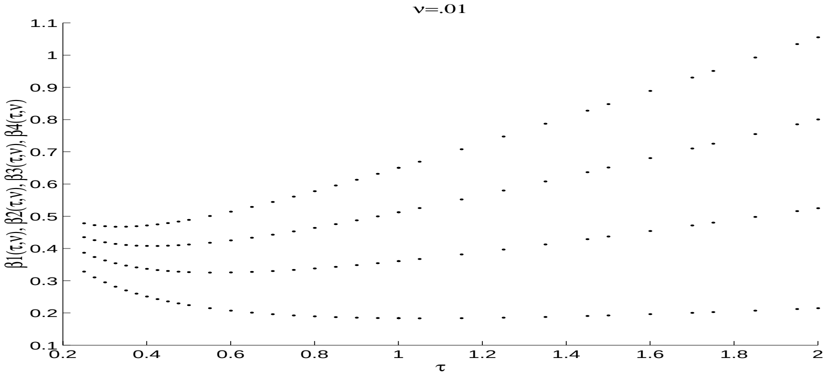


FIG. 5.15. $\beta_j(t,\nu)$ *vs.* t *for* $t \in [.25, 2]$ *and* $j = 1,\dots,4$. $\nu = 10^{-2}$ *and* $N = 5\times 10^4$ *poles. Same parameters as in Fig.* 5.14. *Turn around times at* $t_u \approx 1.05$, $t_u \approx .55$, $t_u \approx .425$, $t_u \approx .325$ *for* $\beta_j(t,\nu), j = 1,\dots,4$.

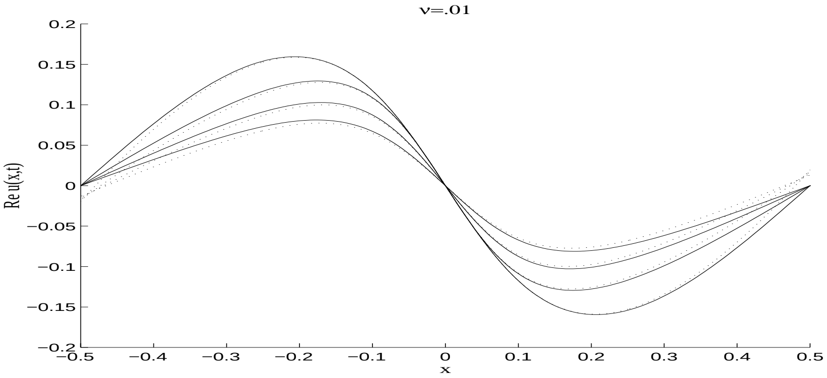


FIG. 5.16. *Comparison of the finite difference approximation (solid) and the pole dynamics (dotted) for* $\nu = 10^{-2}$ *at* $t = .5, 1, 1.5, 2$. *Finite difference mesh size:* $N_x = 100$ *points,* $\Delta x = .5 * 10^{-2}$*,* $N_t = 2{,}000$ *RK*45 *time steps with* $\Delta t = 10^{-3}$. *Pole dynamics: same as Fig.* 5.14.

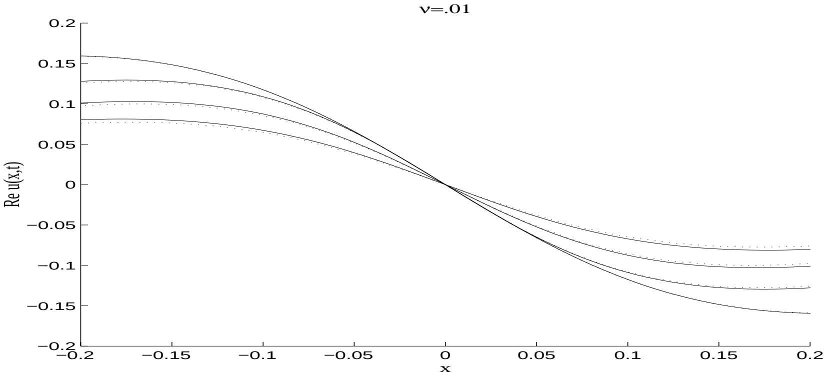


FIG. 5.17. *Closeup of Fig.* 5.6 *in* $[-.2, .2]$.

and the saddle-point approximation at the successive times $t = 1, 1.5, 2$. The mesh size is the same as the one for Figs. 5.2 and 5.3. One can observe that the saddle-point approximation overshoots the true value of the solution which is best captured by the difference scheme. This overshoot is due to the degeneracy of the saddle-point formula at the caustic and the inaccuracies around it. The correct behavior in a neighborhood of this caustic can be correctly described only by the uniform asymptotic expansion of section 3.3.

In Figs. 5.6 and 5.7, we compare the difference method and the pole dynamics for $\nu = 10^{-3}$ with $N = 50{,}000$ poles at the times $t = .5, 1, 1.5, 2$. The pole dynamics are run forward and backward in time starting from $t = t_* = 1$ until $t = .5$ and $t = 2$. The solution is then reconstructed from the pole expansion and the pole locations at these specific times and is compared to the finite difference approximations with mesh size $N_x = 200$ points, $\Delta x = .25 * 10^{-2}$, $N_t = 2{,}000$ RK45 time steps with $\Delta t = 10^{-3}$. The agreement between the finite difference and the pole dynamics close to the shock region is very good as opposed to the tails. Since the pole dynamics simulation involved only $50{,}000$ poles in Fig. 5.6, the mismatch in the tail is characteristic of the slow convergence of the pole expansion in the tails that are displayed in Fig. 5.1 for

$\nu = 10^{-3}$. There is also a small source of error in the difference scheme where the boundary condition at $x = 1/2$ is set to the inviscid value ($u(1/2, t) = 0$). This error in the difference approximation increases for larger ν. Thus, the discrepancy observed in the tails of the solution in Fig. 5.16 is more likely to arise from errors in the difference scheme than the pole dynamics. Indeed, it suffices to look at the convergence of the pole expansion at $t = t_*$ in Fig. 5.1, $\nu = 10^{-2}$, to establish confidence in the pole dynamics.

However, one can notice that regardless of the size of ν (whether $\nu = 10^{-2}$ or 10^{-3}), within the shock region of width $\mathcal{O}(\nu)$, the agreement between the pole dynamics and the difference approximation is very good (see Figs. 5.3, 5.17). This shows that the dynamics of the first few poles is accurately captured by the pole dynamics. This also becomes apparent when comparing the simulations done with varying number of poles (see Figs. 5.10 and 5.11). Finally, it should be noted that increasing the step-size of the time increment (in a reasonable way) in the pole dynamics barely affects the computations (see Figs. 5.8 and 5.9).

We plot the evolution of the first four (ordered) poles on the imaginary axis ($\beta_k, k = 1, \dots, 4$) and focus on the "turn-around" times t_u and the position of the first ordered pole β_1, which determines the width of the analyticity strip. One can see that the behavior of the pole β_1 displayed in Figs. 5.8 and 5.14 is qualitatively similar to the one obtained by Sulem, Sulem, and Frisch in [35, section III-B, Fig. 3] using spectral methods for the initial data $u_0(x) = \sin(x)$ with $\nu = .05$. The most important feature in the behavior of the first ordered pole is clearly the fact that it turns around before crossing the real axis, thus preserving the uniform analyticity of the viscous solution within the strip $|\Im x| \leq \delta_1 < \beta_1$, where $\beta_1(t, \nu) > 0$ for all $t > 0$. Moreover, it is interesting to note that the poles $\pm\beta_k(t, \nu)$ are confined to the imaginary axis and move towards the origin until a time $t = t_u(k)$, $k \in \mathbb{Z}^*$; this is the time at which they turn around and move away from the origin. These turn-around times $t_u(k)$ decrease as k increases: $t_u(1) > t_u(2) > \cdots > t_u(n) > \cdots > 0$. Moreover, $t_u(1)$ occurs before t_* for $\nu \gtrapprox .01$ and after t_* for $\nu \lessapprox .01$. Thus, the last pole to turn around is the first ordered pole β_1, i.e., the one closest to the real axis. For $\nu = 10^{-3}$, the turn-around times for $\beta_j, j = 1, \dots, 4$ are at $t \approx 1.62, 1.51, 1.39, 1.27$, respectively. For $\nu = 10^{-2}$, the turn-around times for $\beta_j, j = 1, \dots, 4$ are at $t \approx 1.05, .55, .425, .325$, respectively. Thus, comparing Figs. 5.13 and 5.15, one can see that the turn-around times $t_u(k)$ increase with decreasing ν. That is, one can relate the time of initial decay of the solution to the turn-around times $t_u(k)$ by comparing the evolution of the poles (see Figs. 5.13, 5.15) to the corresponding evolution of the solution (see Figs. 5.6, 5.16).

Acknowledgment. The author would like to thank the following: Profs. D. Bessis, R. Caflisch, N. Ercolani, J. D. Fournier, L. Greengard, and K. Holczer and Drs. K. Dempsey, L. Cortelezzi, D. Petrasek, H. Rahimizadeh, C. Rippel, and T. Tassa.

REFERENCES

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM, Philadelphia, PA, 1981, pp. 203–209.

[2] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[3] L. V. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.

[4] M. Avellaneda, *Statistics of shocks in Burgers turbulence,* II*: Tail probabilities for velocities, shock-strengths and rarefaction intervals*, Comm. Math. Phys., 169 (1995), pp. 45–59.

[5] C. Bardos and S. Benachour, *Domaine d'analyticité des solutions de l'équation d'Euler dans un ouvert de $\mathbb{R}^n$*, Ann. Scuola Norm. Sup. Pisa, 4 (1977), pp. 647–687.

[6] D. Bessis and J. D. Fournier, *Complex singularities and the Riemann surface for the Burgers equation*, Research Reports in Physics: Nonlinear Physics, Springer-Verlag, Berlin, New York, 1990, pp. 252–257.

[7] D. Bessis and J. D. Fournier, *Pole condensation and the Riemann surface associated with a shock in Burgers equation*, J. Phys. Lett., 45 (1984), pp. L833–L841.

[8] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[9] J. M. Burgers, *The Nonlinear Diffusion Equation*, D. Reidel, Boston, MA, 1974.

[10] J. M. Burgers, *A mathematical model illustrating the theory of turbulence*, Adv. Appl. Mech., 1 (1948), pp. 171–199.

[11] R. E. Caflisch, *A simplified version of the abstract Cauchy–Kowalewski theorem with weak singularities*, Bull. Amer. Math. Soc. (N.S.), 23 (1990), pp. 495–500.

[12] R. E. Caflisch, N. Ercolani, T. Y. Hou, and Y. Landis, *Multi-valued solutions and branch point singularities for nonlinear hyperbolic or elliptic systems*, Comm. Pure Appl. Math., 46 (1993), pp. 453–499.

[13] F. Calogero, *Motion of poles and zeros of special solutions of nonlinear and linear partial differential equations and related "solvable" many body problems*, Nuovo Cimento B (11), 43 (1978), pp. 177–241.

[14] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. and Stat. Comp., 9 (1988), pp. 669–686.

[15] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599–611.

[16] D. V. Choodnovsky and G. V. Choodnovsky, *Pole expansions of nonlinear partial differential equations*, Nuovo Cimento B (11), 40 (1977), pp. 339–353.

[17] J. D. Cole, *On a quasi-linear parabolic equation occurring in aerodynamics*, Quart. Appl. Math., 9 (1951), pp. 225–236.

[18] A. R. Forsyth, *Theory of Differential Equations*, Part IV, Vol. 6, Dover, New York, 1906.

[19] J. D. Fournier and U. Frisch, *L'équation de Burgers deterministe et statistique*, J. Mech. Theory Appl., 2 (1983), pp. 699–750.

[20] U. Frisch and R. Morf, *Intermittency in nonlinear dynamics and singularities at complex times*, Phys. Rev. A (3), 23 (1981), pp. 2673–2705.

[21] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comp. Phys., 73 (1987), pp. 325–348.

[22] E. Hopf, *The partial differential equation $u_t + u\,u_x = \mu\,u_{xx}$*, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[23] N. Joshi and J. A. Petersen, *A method for proving the convergence of the Painlevé expansions of partial differential equations*, Nonlinearity, 7 (1994), pp. 595–602.

[24] D. Kaminski, *Asymptotic expansion of the Pearcey integral near the caustic*, SIAM J. Math. Anal., 20 (1989), pp. 987–1005.

[25] Y. Kimura, *Dynamics of complex singularities for Burgers' equation*, in Proc. NEEDS '94, V. G. Makhankov, ed., World Scientific, Singapore, 1995.

[26] M. D. Kruskal, *The Korteweg–de Vries Equation and Related Evolution Equations*, Lectures in Appl. Math., Amer. Math. Soc., Providence, RI, 1974.

[27] L. Nirenberg, *On the abstract Cauchy–Kowalewski theorem*, J. Differential Geom., 6 (1972), pp. 561–576.

[28] R. B. Paris, *The asymptotic behavior of Pearcey's integral for complex variables*, Proc. Roy. Soc. London Ser. A, 432 (1991), pp. 391–426.

[29] G. Pólya, *Über trigonometrische integrale mit nur reellen nullstellen*, J. Reine Angewandte Math., 158 (1927), pp. 6–18.

[30] Z.-S. She, E. Aurell, and U. Frisch, *The inviscid Burgers equation with initial data of Brownian type*, Comm. Math. Phys., 148 (1992), pp. 623–641.

[31] D. Senouf, *Asymptotic and numerical approximations of the zeros of Fourier integrals*, SIAM J. Math. Anal., 27 (1996), pp. 1102–1128.

[32] D. Senouf, *Dynamics and condensation of complex singularities for Burgers' equation* II, SIAM J. Math. Anal., 28 (1997), pp. 1490–1513.

[33] D. Senouf, R. Caflisch, and N. Ercolani, *Pole dynamics and oscillations for complex Burgers equation in the small dispersion limit*, Nonlinearity, 9 (1996), pp. 1671–1702.

[34] Y. G. Sinai, *Statistics of shocks in solutions of inviscid Burgers equation*, Comm. Math. Phys., 148 (1992), pp. 601–621.

[35] C. Sulem, P. L. Sulem, and H. Frisch, *Tracing complex singularities with spectral methods*,

J. Comp. Phys., 50 (1983), pp. 138–161.

[36] O. Thual, U. Frisch, and M. Hénon, *Application of pole decomposition to an equation governing the dynamics of wrinkled flame fronts*, J. Phys., 46 (1985), pp. 1485–1494.

[37] F. Ursell, *Integrals with a large parameter. Several nearly coincident saddle points*, Proc. Cambridge Philos. Soc., 72 (1972), pp. 49–65.

[38] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley Interscience, New York, 1974.

[39] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.

SIAM J. Math. Anal.
Vol. 28, No. 6, pp. 1490–1513, November 1997

012

DYNAMICS AND CONDENSATION OF COMPLEX SINGULARITIES FOR BURGERS' EQUATION II*

DAVID SENOUF†

Abstract. The zero-viscosity limit of a meromorphic solution to Burgers' equation (BE) is found via an integral representation of the Mittag–Leffler expansion of the solution involving a "polar" measure. The weak zero-viscosity limit of this Borel measure (analogously to the zero-dispersion limit of the spectral measure in the Korteweg–de Vries (KdV) problem) corresponds to the asymptotic density of poles which characterizes their condensation on the imaginary axis. The resulting integral representation of the inviscid solution is computed by residues and is shown to match the characteristic solution up to the inviscid shock time t_*. The continuum limit of the Mittag–Leffler expansion and the Calogero dynamical system (CDS) (which describes the time evolution of the poles) is a system of two integro-differential equations which provide a new representation of the solution to the inviscid BE. For $t \leq t_*$, a uniform asymptotic expansion of the Fourier transform of the inviscid solution is obtained, thereby providing the analyticity properties of the inviscid solution.

Key words. partial differential equations, zero-viscosity limit, pole condensation

AMS subject classifications. 35A20, 35A40, 35B40, 35Q53, 41A60

PII. S0036141095289701

1. Introduction.

In this article we continue the investigation from part I [18] of the spatial analyticity properties of a solution to Burgers' equation (hereafter referred to as "BE"):

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \nu\,\frac{\partial^2 u}{\partial x^2}, \qquad x \in \mathbb{R},\ t > 0,\ \nu \geq 0. \tag{1.1}$$

Previous work concerning the analyticity properties of BE can be found in [3, 4, 13, 14, 18, 20].

We focus on a particular initial value problem (IVP) for (1.1) which was introduced by Fournier and Frisch [13] and further studied by Bessis and Fournier in [3, 4]. In this problem, the initial condition is given by

$$u(x,0) = u_0(x) = 4x^3 - x/t_*, \qquad x \in \mathbb{R}, \tag{1.2}$$

where t_* is a fixed positive parameter. This initial value is chosen for its generic property, which is due to the type of singularity occurring in the inviscid solution ($\nu = 0$) at the shock time $t = -\left(\inf_x u_0'(x)\right)^{-1} = t_*$ (cf. Appendix A for more details).

It was shown in part I that all meromorphic solutions to BE with an odd initial data must have a symmetric pole expansion of the form

$$u_\nu(x,t) = \frac{x}{t} - 2\nu \sum_{n\in\mathcal{I}} \frac{1}{x - a_n(t,\nu)}, \tag{1.3}$$

[*] Received by the editors July 31, 1995; accepted for publication (in revised form) June 25, 1996. This research was partially supported by NSF grant DMS-9306720.

http://www.siam.org/journals/sima/28-6/28970.html

[†] Lehman Brothers, One Broadgate, London EC2M 7HA, United Kingdom (dsenouf@lehman.com).

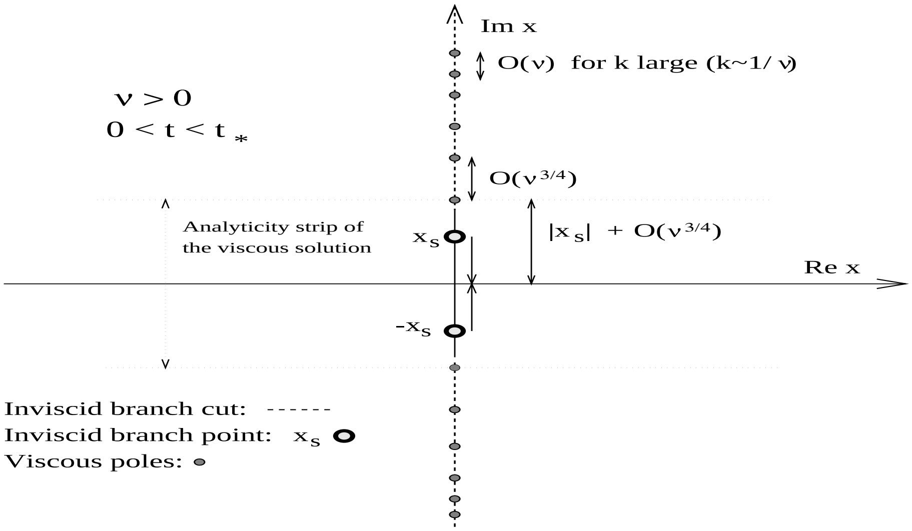


FIG. 1.1. *Inviscid branch points, branch cuts, and viscous poles for $\nu > 0$ and $0 < t < t_*$. The poles are located above the inviscid branch-point singularities according to the asymptotic formula $\beta_k(t, \nu) = \Im x_s(t) + \mathcal{O}(\nu^{3/4})$ as $\nu \to 0^+$ for k fixed. The distance separating two successive poles is asymptotically given by $\Delta\beta_k = \mathcal{O}(\nu)$ as $\nu \to 0^+$ for k large ($k \sim 1/\nu$).*

where $\mathcal{I} \subseteq \mathbb{Z}$ is a finite or countable symmetric set (i.e., if $a_n \in \mathcal{I}, a_{-n} = -a_n \in \mathcal{I}$). Moreover, the poles $\{a_n(t, \nu)\}_{n \in \mathcal{I}}$ must satisfy a Calogero-type dynamical system [8] (hereafter referred to as "CDS") of the form

$$\frac{da_n}{dt} = \frac{a_n}{t} - 2\nu \sum_{\substack{l \in \mathcal{I} \\ l \neq n}} \frac{1}{a_n - a_l} \qquad \forall n \in \mathcal{I} \subseteq \mathbb{Z}. \tag{1.4}$$

The solution to the IVP (1.1)–(1.2) is the meromorphic function

$$u_\nu(x, t) = \frac{x}{t} - 2\nu \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{1}{x - i\beta_n(t, \nu)} = \frac{x}{t} - 2\nu \sum_{n=1}^{\infty} \frac{2x}{x^2 + \beta_n^2(t, \nu)}, \tag{1.5}$$

where $\{\pm i\beta_n\}_{n \in \mathbb{Z}^*}$ is a countable set of pure imaginary conjugate poles (the zeros of the Cole–Hopf variable) satisfying $\beta_{-n} = -\beta_n$. The motion of these poles on the imaginary axis is governed by an infinite-dimensional CDS:

$$\dot{\beta}_n = \frac{\beta_n}{t} - 2\nu \sum_{\substack{l=-\infty \\ l \neq n,0}}^{\infty} \frac{1}{\beta_l - \beta_n} = \frac{\beta_n}{t} + \frac{\nu}{\beta_n} - 2\nu \sum_{\substack{l=1 \\ l \neq n}}^{\infty} \frac{2\beta_n}{\beta_l^2 - \beta_n^2} \quad \forall n \in \mathbb{Z}^*. \tag{1.6}$$

Numerical simulations of the evolution of these poles and the solution for small viscosity are described in [18]. For more details on the derivation of (1.5) and (1.6), see the companion article [18, section 2]. As $\nu \to 0^+$, these poles condense on the imaginary axis for all $t > 0$. The asymptotic distance between two successive poles as $\nu \to 0^+$ is proportional to ν when the index of these poles grows like $k \sim 1/\nu$ (see Fig. 1.1). This condensation phenomenon is captured by an asymptotic density of poles (also referred to as the limiting pole density in the work of Bessis and Fournier

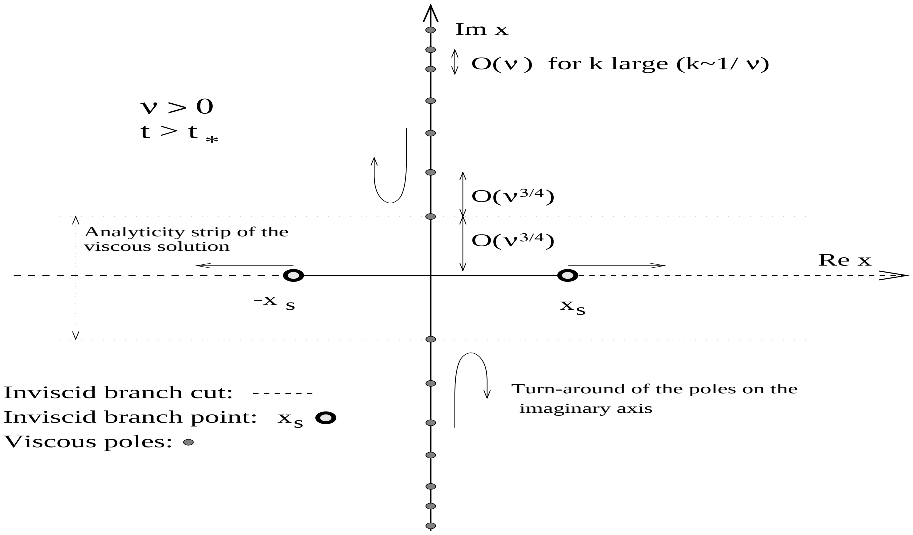

Im x
O(ν) for k large (k~1/ν)
ν > 0
t > t *
O(ν3/4)
O(ν3/4)
Analyticity strip of the viscous solution
Re x
-x s
x s
Inviscid branch cut: ------
Inviscid branch point: x s
Viscous poles:
Turn-around of the poles on the imaginary axis

FIG. 1.2. *Inviscid branch points, branch cuts, and viscous poles for $\nu > 0$ and $t > t_*$. The inviscid branch points have coalesced at $t = t_*$ at the origin and are now moving away from each other on the real axis. However, the poles are fixed to the imaginary axis and are asymptotically given by $\beta_k(t, \nu) = \mathcal{O}(\nu^{3/4})$ as $\nu \to 0^+$ and $\Delta\beta_k = \mathcal{O}(\nu)$ as $\nu \to 0^+$ for k large ($k \sim 1/\nu$). They turn around and move away from the origin at a time $t_u > t_*$ if ν is small enough (if $\nu \gtrapprox .01$, $t_u < t_*$).*

[4]). We show that this density depends directly on the relevant saddle points of the small ν asymptotic expansion of the component $E_\nu(x,t)$ which carries the zeros of the Cole–Hopf variable. From the Cole–Hopf transformation, we find that

$$u_\nu(x,t) = \frac{x}{t} - 2\nu\,\partial_x \log\bigl(E_\nu(x,t)\bigr), \quad E_\nu(x,t) = \int_{-\infty}^{\infty} e^{w(z,x)/2\nu}\,dz,$$

where $w(z,x)$ is the phase function defined by

$$w(z,x) = \int_0^z \left(\frac{x}{t} - \frac{\eta}{t} - u_0(\eta)\right) d\eta.$$

The saddle points $z_s(\beta,t)$ of the phase function $w(z,x)$ are implicitly given by

$$0 = w_z(z_s, i\beta) = \frac{i\beta}{t} - \frac{z_s}{t} - u_0(z_s).$$

Let $\sigma(\beta;t)$ be a cumulative distribution function corresponding to the integral of the asymptotic density of poles (which we will define later). That is, $\sigma(\beta;t)$ counts the number of poles contained within the interval $[0,\beta]$. Then it is shown that

$$\sigma(\beta;t) = \frac{\Im w(z_s, i\beta)}{\pi},$$

and then the asymptotic pole locations are implicitly given by the equation

$$\frac{\sigma(\beta;t)}{2\nu} = k - 1/2, \qquad k \to +\infty.$$

We show that the asymptotic density of poles defined by Bessis and Fournier in [3, 4] as

$$\rho(\beta;t) = \lim_{\substack{n\to\infty \\ \nu\to 0^+}} \left.\frac{2\nu}{\Delta\beta_n(t,\nu)}\right|_{\beta_n=\beta}$$

is given by

$$\rho(\beta;t) = \frac{d}{d\beta}\sigma(\beta;t) = \frac{1}{\pi t}\,\Re z_s^+(\beta;t),$$

where $z_s^+(\beta;t)$ is the relevant saddle point with a positive real part in the expansion of $E_\nu(i\beta,t)$ as $\nu \to 0^+$. This density is explicitly calculated for all $t > 0$ using Cardan's formula. The Mittag–Leffler expansion (1.5) has an integral representation which is valid away from the imaginary axis. It can be expressed as the integral of a continuous function against the distributional derivative of a nonnegative regular finite Borel measure defined for $|\beta| \le \beta_{\max} < +\infty$ as

$$\sigma_\nu(\beta;\beta_{\max},t) = \int_{-\beta_{\max}}^{\beta} 2\nu \sum_{k=1}^{N_\nu} \left[\delta\left(\xi - \beta_k(t,\nu)\right) + \delta\left(\xi + \beta_k(t,\nu)\right)\right] d\xi,$$

where

$$N_\nu(\beta_{\max}) = \sup_{0<\delta\le t\le t_*} \left[\frac{\sigma(\beta_{\max};t)}{2\nu}\right] < +\infty. \tag{1.7}$$

The "polar" measure $\sigma_\nu(\beta;\beta_{\max},t)$ is analogous to the spectral measure in the KdV problem (cf. [12, 15]). The zero-viscosity limit of the pole expansion is found by taking the weak limit of $d\sigma_\nu(\beta;\beta_{\max},t)/d\beta$ which approximates the asymptotic density of poles. Thus, we show that for $\beta \in [-\beta_{\max},\beta_{\max}]$, $0 < \beta_{\max} < +\infty$,

$$\rho(\beta;t) \equiv \text{w-}\lim_{\nu\to 0^+} \frac{d\sigma_\nu}{d\beta}(\beta;\beta_{\max},t),$$

where w-$\lim_{\nu\to 0^+}$ denotes the weak limit of measures. The limiting integral representation of the solution is given by the nonparametric form

$$u(x,t) = \frac{x}{t} - \int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{x - i\beta}\, d\beta = \frac{x}{t} - x\int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{x^2+\beta^2}\, d\beta.$$

This representation is computed via residues, and the analytic structure of the inviscid solution is explicitly recovered up to t_*.

We also show that the continuum limit of the pair of equations consisting of the pole expansion (1.5) and the dynamical system (1.6) is the system of two integro-differential equations in parametric form:

$$\begin{aligned}\frac{\partial f}{\partial t}(\zeta,t) &= \frac{f(\zeta,t)}{t} - P.V.\int_{-\infty}^{\infty} \frac{d\zeta'}{f(\zeta,t) - f(\zeta',t)} \\ &= \frac{f(\zeta,t)}{t} - f(\zeta,t)\, P.V.\int_{-\infty}^{\infty} \frac{d\zeta'}{f^2(\zeta,t) - f^2(\zeta',t)}\end{aligned}$$

and

$$u(x,t) = \frac{x}{t} - \int_{-\infty}^{\infty} \frac{d\zeta'}{x - f(\zeta',t)} = \frac{x}{t} - x\int_{-\infty}^{\infty} \frac{d\zeta'}{x^2 - f^2(\zeta',t)},$$

in which the spatial branch cuts are defined by the condition $x \neq f(\zeta, t)$ for each fixed $t > 0$. This 2×2 system is shown to be equivalent to the characteristic equations of the inviscid BE.

The equivalence between the parametric and nonparametric form of the integral representation of the inviscid solution is obtained by introducing a simple change of variable $i\beta = f(\zeta, t)$ in the parametric equations. From this analysis, we clarify the relation between the pole positions and their asymptotic density. We show for $t = t_*$ that the pole positions can be recovered from the asymptotic density by choosing the right discretization on the "continuum" curve on which this density lies.

Furthermore, the analyticity properties of the inviscid solution can be analyzed by describing the asymptotic behavior of its Fourier transform (see [13, 20]). We find a uniform asymptotic expansion as $k \to +\infty$ of the Fourier transform of the inviscid solution, clarifying the seemingly discontinuous change of behavior of $\hat{u}(k,t)$ at t_* presented in [13]. This discontinuity in the asymptotic behavior is a direct consequence of the coalescence of the two second-order branch points $\pm x_s(t)$ into a third-order branch point at the origin $x_s(t_*) = 0$. We show that as $k \to +\infty$, $\hat{u}(k,t) = \mathcal{C}_0 \cdot (tk)^{-4/3}\mathrm{Ai}[(-3ikx_s(t)/2)^{2/3}]\left(1 + \mathcal{O}\left(k^{-1}\right)\right)$. From the (uniform) asymptotic expansion of the Airy function we find that $\hat{u}(k,t) \sim \mathcal{C}_1(t) \cdot (t_* - t)^{-1/4}\, k^{-3/2} \exp(-k|x_s(t)|)$ for $0 < t < t_*$ and $\hat{u}(k, t_*) \sim \mathcal{C}_2 \cdot (t_* k)^{-4/3}$, where $\mathcal{C}_0, \mathcal{C}_1(t), \mathcal{C}_2$ are appropriate numerical constants.

2. Polar measure, integral representation, and inviscid limit.

In [18], the solution to the IVP (1.1)–(1.2) is constructed in the following property.

PROPERTY 2.1. *Let $2\alpha = 1/t_* - 1/t$ for $\nu, t, t_* > 0$; then*

$$\begin{aligned} u_\nu(x,t) &= \frac{x}{t} - 2\nu\, \partial_x \log\bigl(E_\nu(x,t)\bigr), \\ E_\nu(x,t) &= \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\nu}\left(\frac{x}{t}y + \alpha y^2 - y^4\right)\right\} dy. \end{aligned}$$

Furthermore, $u_\nu(x,t)$ has a Mittag–Leffler (pole) representation:

$$u_\nu(x,t) = \frac{x}{t} - 2\nu \sum_{n=1}^{\infty} \frac{2x}{x^2 + \beta_n^2(t,\nu)},$$

which converges uniformly on compact sets for x away from the poles $x = \pm i\beta_n$.

From the integral representation of the solution one can describe the behavior of the solution as the viscosity tends to zero: using a saddle-point analysis, we have shown in [18, section 4.2] that the dominant behavior of $E_\nu(i\beta, t)$ as $\nu \to 0^+$ or as $\beta \to +\infty$ is given by an asymptotic relation of the form

$$\begin{aligned} &\sqrt{\frac{|6z_0(\beta;t)^2 - \alpha|}{2\pi\nu}}\ \exp\left\{-\frac{1}{2\nu}\Re w\bigl(z_0(\beta;t), i\beta\bigr)\right\} E_\nu(i\beta,t) \\ &\qquad = \cos\left(\frac{\Im w(z_0(\beta;t), i\beta)}{2\nu} - \frac{\theta(z_0(\beta;t),t)}{2}\right) + \mathcal{O}\left(\frac{\nu}{\beta^{4/3}}\right), \end{aligned}$$

where

$$\begin{aligned} w(z, i\beta) &= \int_0^z (i\beta/t - \eta/t - u_0(\eta))\, d\eta = i\beta z/t + \alpha z^2 - z^4, \\ \theta(z,t) &= \arg(\partial_z^2 w) = \arg(6z^2 - \alpha), \quad -\pi \le \theta(z,t) < \pi, \end{aligned}$$

and $z_s(\beta;t)$ is implicitly defined by

$$0 = w_z(z_s(\beta;t), i\beta) = \frac{i\beta}{4t} + \frac{\alpha}{2}z_s - z_s^3, \qquad s = 0, 1, 2. \tag{2.1}$$

Solving (2.1) using Cardan's formula (see Appendix B) and separating real and imaginary parts, we find that

$$\begin{cases} z_0 = \frac{\sqrt{3}}{2}(\mathcal{A} - \mathcal{B}) + \frac{i}{2}(\mathcal{A} + \mathcal{B}), \\ z_1 = -\overline{z_0} = \frac{\sqrt{3}}{2}(\mathcal{B} - \mathcal{A}) + \frac{i}{2}(\mathcal{A} + \mathcal{B}), \\ z_2 = -i(\mathcal{A} + \mathcal{B}), \end{cases} \tag{2.2}$$

where for $\beta > |x_s(t)|$,

$$\begin{cases} \mathcal{A}(\beta;t) = (8t)^{-1/3} \sqrt[3]{\beta + \sqrt{\beta^2 + x_s^2}} > 0, \\ \mathcal{B}(\beta;t) = (8t)^{-1/3} \sqrt[3]{\beta - \sqrt{\beta^2 + x_s^2}} > 0. \end{cases} \tag{2.3}$$

For $\beta < -|x_s(t)|$, $\mathcal{A}$ and $\mathcal{B}$ are defined by the odd parity condition $\mathcal{A}(-\beta;t) = -\mathcal{A}(\beta;t)$ and $\mathcal{B}(-\beta;t) = -\mathcal{B}(\beta;t)$ so that

$$z_s(-\beta;t) = -z_s(\beta;t). \tag{2.4}$$

For small ν (fixed k) or for large β (large k, fixed ν), the poles β_k are approximated by the roots of the equation

$$\frac{\sigma(\beta;t)}{2\nu} - \frac{1}{2\pi}\theta(z_0(\beta;t), t) = k - \frac{1}{2}, \quad k \in \mathbb{N}^*, \tag{2.5}$$

with the convention that $\beta_{-k} \equiv -\beta_k$. Clearly, since $|\theta| \leq \pi$, the contribution of θ is negligible compared with that of σ. Thus, we approximate (2.5) by

$$\frac{\sigma(\beta;t)}{2\nu} \approx k - \frac{1}{2}, \quad k \in \mathbb{N}^*. \tag{2.6}$$

Choose a parameter $\beta_{\max} < +\infty$, and then choose $N_\nu(\beta_{\max})$ as follows: let Int$[x]$ denote the integer part of x with half-integers rounded down, and fix $\nu > 0$. Then for any $\delta > 0$ and compact set $[\delta, t_*]$, define N_ν by

$$N_\nu(\beta_{\max}) = \sup_{0<\delta \leq t \leq t_*} \left[\frac{\sigma(\beta_{\max};t)}{2\nu}\right] < +\infty. \tag{2.7}$$

The $\beta_k(t,\nu)$ are ordered as follows:

$$0 \leq |x_s(t)| < \beta_1(t,\nu) < \cdots < \beta_k(t,\nu) < \cdots < \beta_{N_\nu} < +\infty \tag{2.8}$$

for $1 < k < N_\nu$. For negative indices, the ordering of the β_{-k}'s is the reverse of that given in (2.8).

Let $U_\nu^{\beta_{\max}}(x,t)$ be the N_νth partial sum of $U_\nu(x,t)$:

$$\begin{aligned} U_\nu^{\beta_{\max}}(x,t) = x - t\, u_\nu^{\beta_{\max}}(x,t) &= t \cdot 2x \cdot \sum_{n=1}^{N_\nu} \frac{2\nu}{x^2 + \beta_n^2} \\ &= t \cdot 2\nu \sum_{n=1}^{N_\nu} \left(\frac{1}{x - i\beta_n} + \frac{1}{x + i\beta_n}\right). \end{aligned} \tag{2.9}$$

Let $U_\nu(x,t)$ be the spatially singular part of the viscous solution defined by

$$U_\nu(x,t) = x - t\,u_\nu(x,t) = t \cdot 2x \sum_{n=1}^{\infty} \frac{2\nu}{x^2 + \beta_n^2(t,\nu)}, \tag{2.10}$$

and let the remainder $R_\nu^{\beta_{\max}}(x,t)$ be defined by

$$R_\nu^{\beta_{\max}}(x,t) = U_\nu(x,t) - U_\nu^{\beta_{\max}}(x,t) = t \cdot 2x \sum_{n=N_\nu+1}^{\infty} \frac{2\nu}{x^2 + \beta_n^2(t,\nu)}.$$

Let $\delta(\beta)$ denote the usual Dirac measure, and define the density $\sigma_\nu(\beta;\beta_{\max},t)$ with support in $[-\beta_{\max},\beta_{\max}]$ by

$$\sigma_\nu(\beta;\beta_{\max},t) = \int_{-\beta_{\max}}^{\beta} 2\nu \sum_{k=1}^{N_\nu} \left[\delta\left(\xi - \beta_k(t,\nu)\right) + \delta\left(\xi + \beta_k(t,\nu)\right)\right] d\xi. \tag{2.11}$$

Since the poles β_n are ordered according to (2.8), this insures that (2.11) is nonnegative and vanishes outside $[-\beta_{\max},\beta_{\max}]$. Moreover, once $\beta_{\max}$ has been chosen, $\sigma_\nu(\beta;\beta_{\max},t)$ is uniformly bounded for $0 < \delta \le t \le t_*$ by $2\nu N_\nu$. Thus, from the definition (2.7) of N_ν, $\sigma_\nu(\beta;\beta_{\max},t)$ is uniformly bounded in ν, and thus is a regular finite Borel measure. Since

$$\begin{aligned} d\sigma_\nu(\beta;\beta_{\max},t) &= 2\nu \sum_{k=1}^{N_\nu} \left[\delta\left(\beta - \beta_k(t,\nu)\right) + \delta\left(\beta + \beta_k(t,\nu)\right)\right] d\beta \\ &= 2\nu \sum_{\substack{k=-N_\nu \\ k\neq 0}}^{N_\nu} \delta\left(\beta - \beta_k(t,\nu)\right) d\beta, \end{aligned} \tag{2.12}$$

the measure $d\sigma_\nu(\beta;\beta_{\max},t)$ consists of a sum of delta functions with weight (height) 2ν decreasing as $\nu \to 0^+$. The density of these delta functions increases like $1/\Delta\beta_k = \mathcal{O}(1/\nu)$ as $\nu \to 0^+$ for $k \sim 1/\nu$. Since the measure $d\sigma_\nu$ is odd ($d\sigma_\nu(-\beta;\beta_{\max},t) = -d\sigma_\nu(\beta;\beta_{\max},t)$), we can represent $U_\nu^{\beta_{\max}}(x,t)$ as

$$U_\nu^{\beta_{\max}}(x,t) = t \cdot \int_{-\infty}^{\infty} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x - i\beta} = t \cdot 2x \int_0^{\infty} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x^2+\beta^2}, \tag{2.13}$$

where the last integral should be understood as

$$\int_0^{\infty} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x^2+\beta^2} \equiv \frac{1}{2}\int_{-\infty}^{\infty} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x^2+\beta^2}. \tag{2.14}$$

Since the measure $d\sigma_\nu(\beta;\beta_{\max},t)$ has support in the compact interval $[-\beta_{\max},\beta_{\max}]$, we have proved the following property.

Property 2.2. *$U_\nu^{\beta_{\max}}(x,t)$ has an integral representation for $x \notin [-i\beta_{\max}, i\beta_{\max}]$ given by*

$$U_\nu^{\beta_{\max}}(x,t) = t \cdot \int_{-\beta_{\max}}^{\beta_{\max}} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x - i\beta} = t \cdot x \int_{-\beta_{\max}}^{\beta_{\max}} \frac{d\sigma_\nu(\beta;\beta_{\max},t)}{x^2+\beta^2}.$$

Now that we have established the validity of the integral representation of Property 2.2, we use this to derive an integral representation of the inviscid solution: we introduce the asymptotic pole density $\rho(\beta;t)$, which also corresponds to the asymptotic distribution of the zeros of E_ν. We define it in relation to the limiting measure $\sigma(\beta;t)$ as follows.

DEFINITION 2.1. *For $\beta \in [-\beta_{\max}, \beta_{\max}]$, the cumulative distribution function $\sigma(\beta;t)$ represents the number of poles within the interval $[0,\beta] \subset [0,\beta_{\max}]$.*

$$\sigma(\beta;t) = \int_0^\beta \rho(\zeta;t)\,d\zeta = \frac{1}{\pi}\Im w(z_s(\beta;t), i\beta).$$

It follows directly from the definition (2.5) of the zeros $\pm i\beta_k(t,\nu)$ of $E_\nu(x,t)$ and from Definition 2.1 of the asymptotic density of the zeros that

$$\lim_{\nu\to 0^+} \sum_{k=1}^{N_\nu} \frac{2\nu}{x^2+\beta_k^2(t,\nu)} = \int_{-\beta_{\max}}^{\beta_{\max}} \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta = \frac{U^{\beta_{\max}}(x,t)}{tx} \tag{2.15}$$

for a fixed $\delta > 0$ and $0 < \delta \le t \le t_*$. Thus, we have that

$$|U^{\beta_{\max}}(x,t) - U_\nu^{\beta_{\max}}(x,t)| = t\,|x| \cdot \left| \int_{-\beta_{\max}}^{\beta_{\max}} \frac{d\sigma - d\sigma_\nu}{x^2+\beta^2} \right| < \epsilon/3 \tag{2.16}$$

for ν small enough on compact sets for x and t away from the branch cuts defined by $(-i\infty, -i|x_s|] \cup [i|x_s|, +i\infty)$ (for a similar argument see, for example, [12]). Thus the convergence of the measure $d\sigma_\nu$ to $d\sigma$ is described in the following way.

PROPERTY 2.3. *For $\beta \in [-\beta_{\max}, \beta_{\max}]$ and $0 < \delta \le t \le t_*$, the sequence of distributions $d\sigma_\nu(\beta;\beta_{\max},t)$ converges weakly to $d\sigma(\beta;t)$:*

$$\text{w-}\lim_{\nu\to 0^+} d\sigma_\nu(\beta;\beta_{\max},t) = d\sigma(\beta;t) = \rho(\beta;t)\,d\beta = \frac{1}{\pi}\Im dw(z_s(\beta;t), i\beta).$$

This measure is analogous to the spectral measure introduced in [12, 15]. Here w-$\lim_{\nu\to 0^+}$ stands for a limit in the sense of weak convergence of measures; that is, w-$\lim_{\nu\to 0^+} d\mu_\nu(\beta) = d\mu(\beta)$ if

$$\lim_{\nu\to 0^+} (\phi, d\mu_\nu) = (\phi, d\mu) = \int_{-\beta_{\max}}^{\beta_{\max}} \phi(\beta)\,d\mu(\beta)$$

for every continuous function ϕ in $[-\beta_{\max}, \beta_{\max}]$. Note that we suspect the convergence

$$\lim_{\nu\to 0^+} U_\nu^{\beta_{\max}}(x,t) = U^{\beta_{\max}}(x,t) \tag{2.17}$$

to hold uniformly over compact sets for t and x away from the branch cuts.

From the definition of the limiting function $U(x,t)$

$$U(x,t) = t \cdot x \int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta,$$

the remainder $R^{\beta_{\max}}(x,t)$, defined for $x \notin (-i\infty, -i\beta_{\max}] \cup [i\beta_{\max}, i\infty)$ as

$$R^{\beta_{\max}}(x,t) = U^{\beta_{\max}}(x,t) - U(x,t) = t \cdot 2x \cdot \int_{|\beta|\ge\beta_{\max}} \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta, \tag{2.18}$$

can be shown to go to zero as $\beta_{\max} \to +\infty$ independently of ν: fix an $R > 0$ such that $|x| \leq R < \beta_{\max}$, then $|x^2 + \beta^2| \geq \beta^2 - R^2$. Let $\theta > 1$ be a fixed parameter; then for $\beta > \beta_{\max} > \sqrt{\theta/(\theta-1)}R$, we have $1/(\beta^2 - R^2) < \theta/\beta^2$. Then since $\rho(\beta; t) = \mathcal{O}(\beta^{1/3})$ as $\beta \to +\infty$ (see Theorem 3.1), we can estimate (2.18) as $\beta_{\max} \to +\infty$ as follows:

$$\begin{aligned}\left|\int_{|\beta|\geq\beta_{\max}} \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta\right| &\leq \mathcal{C}(t)\int_{|\beta|\geq\beta_{\max}} \frac{\beta^{1/3}}{\beta^2 - R^2}\,d\beta \\ &\leq \theta\cdot\mathcal{C}(t)\cdot\int_{|\beta|\geq\beta_{\max}} \frac{d\beta}{\beta^{5/3}} = \mathcal{O}(\beta_{\max}^{-2/3}).\end{aligned}$$

Therefore, on compact sets for x and t,

$$|U^{\beta_{\max}}(x,t) - U(x,t)| < \epsilon/3 \tag{2.19}$$

for $\beta_{\max}$ large enough independent of ν.

The last estimate concerns

$$\begin{aligned}|R_\nu^{\beta_{\max}}(x,t)| &= |U_\nu^{\beta_{\max}}(x,t) - U_\nu(x,t)| \\ &= t\,|2x|\left|\sum_{n=N_\nu+1}^{\infty} \frac{2\nu}{x^2+\beta_n^2(t,\nu)}\right|.\end{aligned} \tag{2.20}$$

In [18, section 4.2], it is shown that $\beta_n(t,\nu) = \mathcal{O}\left((n\nu)^{3/4}\right)$ as $n \to +\infty$ for fixed $\nu > 0$. Therefore, let $y_n(t,\nu) = \mathcal{C}n\nu$, where $\mathcal{C}$ is an appropriate asymptotic constant which depends on t (see (5.1) for such a representation). This assumption also can be justified by combining (5.1) and the fact that the order $\lambda = 4/3$ of the entire function E_ν is also the order of convergence of its zeros (see [18, section 2.1]). Then since $N_\nu(\beta_{\max}) = \sup_{0<\delta\leq t\leq t_*} \operatorname{Int}[\sigma(\beta_{\max};t)/2\nu]$, following a similar argument as in the proof of (2.19), we may estimate (2.20) as

$$\begin{aligned}|U_\nu^{\beta_{\max}}(x,t) - U_\nu(x,t)| &\leq t\,|2x|\left|\sum_{n=N_\nu+1}^{\infty} \frac{2\nu}{x^2+y_n^{3/2}}\right| \\ &\leq t\,|2x|\,\mathcal{C}_1\left|\int_{\beta_{\max}}^{+\infty} \frac{dy}{x^2+y^{3/2}}\right| = \mathcal{O}(\beta_{\max}^{-1/2})\end{aligned}$$

as $\beta_{\max} \to +\infty$, uniform in ν on compact sets for $t \in [\delta, t_*]$ and x away from the branch cuts. Choosing $\beta_{\max}$ large enough so that $|U(x,t) - U^{\beta_{\max}}(x,t)| < \epsilon/3$ and $|U_\nu^{\beta_{\max}}(x,t) - U_\nu(x,t)| < \epsilon/3$ independent of ν and then choosing ν small enough in such a way that $|U^{\beta_{\max}}(x,t) - U_\nu^{\beta_{\max}}(x,t)| < \epsilon/3$, we finally have

$$\begin{aligned}|U_\nu(x,t) - U(x,t)| \leq{}& |U(x,t) - U^{\beta_{\max}}(x,t)| \\ &+ |U^{\beta_{\max}}(x,t) - U_\nu^{\beta_{\max}}(x,t)| + |U_\nu^{\beta_{\max}}(x,t) - U_\nu(x,t)| < \epsilon.\end{aligned}$$

Using the fact that $\rho(\beta; t) = 0$ for $|x| < |x_s(t)|$ when $0 < t \leq t_*$, we have proved the following theorem.

THEOREM 2.4. *For $\delta > 0$, $t \in [\delta, t_*]$; on compact sets for x away from the branch cuts defined by $(-i\infty, -i|x_s|] \cup [i|x_s|, +i\infty)$, we have*

$$\lim_{\nu\to 0^+} U_\nu(x,t) = t\cdot 2x\int_{|x_s(t)|}^{\infty} \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta.$$

3. Asymptotic density of poles. Now that we have defined the asymptotic density of poles, we proceed with its explicit computation for the different time intervals $(0, t_*)$, $t = t_*$ and $(t_*, +\infty)$. As in section 2, let w-$\lim_{\nu\to 0^+}$ denote a weak limit in the sense of weak convergence of measures. Then we prove the following.

THEOREM 3.1. it For $0 < \beta_{\max} < +\infty$, *the asymptotic density of poles* $\rho(\beta; t)$: $[-\beta_{\max}, \beta_{\max}] \times \mathbb{R}_+ \to \mathbb{R}_+$ *is a positive even function of* β *defined by*

$$\rho(\beta;t) \equiv \text{w-}\lim_{\nu\to 0^+} \frac{d\sigma_\nu}{d\beta}(\beta;\beta_{\max},t) = \frac{d\sigma}{d\beta}(\beta;t) = \frac{1}{\pi t}\,\Re z_s^+(\beta;t),$$

where $z_s^+(\beta;t)$ *is the saddle point with positive real part which is relevant to the asymptotic expansion of* $E_\nu(i\beta, t)$ *as* $\nu \to 0^+$. *This saddle point is determined by the implicit equation*

$$\frac{\partial w}{\partial z}\big(z_s(\beta;t), i\beta\big) = \frac{i\beta}{t} - \frac{z_s(\beta;t)}{t} - u_0\big(z_s(\beta;t)\big) = 0.$$

Let $\pm x_s(t) = \pm i\,(3t_*)^{-3/2}(t_* - t)^{3/2}t^{-1/2}$ *be the second-order branch points of the inviscid solution arising from the initial data* $u_0(x) = 4x^3 - x/t_*$. *For* $t > t_*$ *and for* $t < t_*$, $|\beta| > |x_s|$,

$$\rho(\beta;t) = \frac{2^{2/3}\sqrt{3}}{\pi}(4t)^{-4/3}\left\{\sqrt[3]{|\beta| + \sqrt{\beta^2 + x_s^2}} - \sqrt[3]{|\beta| - \sqrt{\beta^2 + x_s^2}}\right\}.$$

For $t < t_*$, $|\beta| < |x_s|$,

$$\rho(\beta;t) = 0.$$

For $t = t_*$,

$$\rho(\beta;t_*) = \frac{2\sqrt{3}}{\pi}(4t_*)^{-4/3}|\beta|^{1/3}.$$

For $t > t_*$, $\beta = 0$,

$$\rho(0;t) = \lim_{\substack{\beta\to 0\\ t>t_*}} \rho(\beta;t) = \frac{1}{2\pi}(t - t_*)^{1/2}t^{-3/2}t_*^{-1/2}.$$

Proof. In [3, 4], Bessis and Fournier introduced a limiting density of poles which characterizes the process of condensation of poles on the imaginary axis as the viscosity $\nu \to 0^+$. They defined it in [3] in the following way:

$$\rho(\beta;t) \equiv \lim_{\substack{n\to\infty\\ \nu\to 0^+}} \left.\frac{2\nu}{\Delta\beta_n(t,\nu)}\right|_{\beta_n=\beta} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+,$$

with $\Delta\beta_n(t,\nu) = \beta_{n+1}(t,\nu) - \beta_n(t,\nu) > 0$, where $n \in \mathbb{Z}\backslash\{0\} = \{\pm 1, \pm 2, \dots\}$, with the convention that $\beta_{-n} = -\beta_n$. Since

$$\frac{\partial w}{\partial \beta}\big(z_s(\beta;t), i\beta\big) = \frac{i}{t}\,z_s(\beta;t), \tag{3.1a}$$

$$\frac{\partial w}{\partial z}\big(z_s(\beta;t), i\beta\big) = \frac{i\beta}{t} - \frac{z_s(\beta;t)}{t} - u_0\big(z_s(\beta;t)\big) = 0. \tag{3.1b}$$

From Property 2.3 we find that the density is given by

$$\rho(\beta;t) = \frac{1}{\pi}\Im\frac{dw}{d\beta}\big(z_s(\beta;t), i\beta\big) = \frac{1}{\pi t}\,\Re z_s^+(\beta;t), \tag{3.2}$$

where $z_s^+(\beta;t)$ is the relevant saddle point with positive real part. Since $z_s(-\beta;t) = -z_s(\beta;t)$, in order to have $\rho(-\beta;t) = \rho(\beta;t) > 0$, we must take in both cases ($\beta > 0$ and $\beta < 0$) the saddle point with positive real part. That is, we must take z_0 for $\beta > 0$ and z_1 for $\beta < 0$ since they are related by $z_0(-\beta;t) = -z_0(\beta;t) = \overline{z_1}(\beta;t)$ and $z_1(-\beta;t) = -z_1(\beta;t) = \overline{z_0}(\beta;t)$ (see [18, section 4.2]). With this choice of saddle points, the asymptotic density defined in (3.2) is positive whether $\beta > 0$ or $\beta < 0$. Note that if we let $x = i\beta$, $x_0 = x_0(x,t) = z_s(\beta;t)$, then the inviscid solution is $u(x,t) = u_0(x_0(x,t),t)$, where $x - x_0 - t\,u_0(x_0) = 0$ (see Appendix C). Combining (2.2) and (3.2), we immediately have an expression for the density as a function of β and t:

$$\rho(\beta;t) = \frac{\sqrt{3}}{2\pi t}(\mathcal{A}(\beta;t) - \mathcal{B}(\beta;t)), \tag{3.3}$$

with $\mathcal{A}(\beta;t)$ and $\mathcal{B}(\beta;t)$ defined in (2.3).

We can now describe the various cases $t = t_*$, $0 < t < t_*$, $t > t_*$, and $x = 0, t > t_*$:

(i) $t = t_*$:

$$\rho(\beta;t_*) = \frac{1}{\pi t_*}\Re z_s(\beta;t_*) = \frac{\sqrt{3}}{2\pi t_*}\left(\frac{\beta}{4t_*}\right)^{1/3} = \frac{2\sqrt{3}}{\pi}(4t_*)^{-4/3}|\beta|^{1/3}. \tag{3.4}$$

In the last step of (3.4), we replace $\beta^{1/3}$ by $|\beta|^{1/3}$ to allow for both $\beta > 0$ and $\beta < 0$. Note that we can obtain (3.4) by taking the limit as $t \to t_*$ in (3.7) or (3.8). It is interesting to see that this density is the only one which explicitly can be computed from the formula introduced by Bessis and Fournier:

$$\rho(\beta;t) = \lim_{\substack{n\to\infty \\ \nu\to 0^+}} \left.\frac{2\nu}{\Delta\beta_n(t,\nu)}\right|_{\beta_n=\beta}. \tag{3.5}$$

The explicit pole positions $\beta_k(t_*,\nu)$ are given in part I by

$$\beta_k(t_*,\nu) = 4t_*\left(\frac{2\nu}{3\sqrt{3}}\right)^{3/4}\cdot\Big((k-1/3)^{3/4} + \mathcal{O}(1/k^{3/4})\Big) \tag{3.6}$$

as $k \to +\infty$ for all ν. Combining (3.5) with (3.6), one recovers (3.4).

(ii) $0 < t < t_*$: the density is zero for $|\beta| \le |x_s|$ and for $|\beta| > |x_s|$;

$$\rho(\beta;t) = \frac{2^{2/3}\sqrt{3}}{\pi}(4t)^{-4/3}\left\{\sqrt[3]{|\beta| + \sqrt{\beta^2 + x_s^2}} - \sqrt[3]{|\beta| - \sqrt{\beta^2 + x_s^2}}\right\}. \tag{3.7}$$

The behavior of $\rho(\beta;t)$ in a neighborhood of $\beta = \pm|x_s|$, ($|\beta| > |x_s|$) is

$$\rho(\beta;t) = \frac{t^{-4/3}}{\sqrt{6}\pi}\frac{\sqrt{|\beta| - |x_s|}}{|x_s|^{1/6}} + \mathcal{O}\Big(\big(|\beta| - |x_s|\big)^{3/2}\Big),$$

as mentioned in [3].

(iii) $t > t_*$: $\forall \beta \in \mathbb{R}$,

$$\rho(\beta;t) = \frac{2^{2/3}\sqrt{3}}{\pi}(4t)^{-4/3}\left\{\sqrt[3]{\beta+\sqrt{\beta^2+x_s^2}}+\sqrt[3]{-\beta+\sqrt{\beta^2+x_s^2}}\right\}. \tag{3.8}$$

(iv) $x = 0, t \geq t_*$: an interesting case occurs at the origin for $t > t_*$, as was pointed out by Bessis and Fournier in [3, 4]. The inviscid solution $u(0,t)$ at the shock is given by the asymptotic density of poles $\rho(0;t)$ (see (3.12)). If we look at the solution at the origin ($\beta \to 0$), $z_s(0;t)$ is the solution to

$$-\frac{z_s(0;t)}{t} = u_0(z_s(0;t)).$$

When $u_0(x) = 4x^3 - x/t_*$,

$$-\frac{z_s(0;t)}{t} = 4z_s(0;t)^3 - \frac{z_s(0;t)}{t_*}.$$

The nonzero pair of opposite saddle points are, therefore,

$$z_s^{\pm}(0;t) = \pm\frac{1}{2}\sqrt{\frac{t-t_*}{tt_*}} = \pm\sqrt{\frac{\alpha}{2}} \geq 0 \qquad \text{when } \ t \geq t_*.$$

The corresponding density is easily found to be

$$\rho(0;t) = \frac{1}{\pi t}\,\Re z_s^+(0;t) = \begin{cases} \frac{1}{2\pi}(t-t_*)^{1/2}t^{-3/2}t_*^{-1/2} & t > t_*, \\ 0 & t \leq t_*. \end{cases} \tag{3.9}$$

This could have been found by letting $\beta \to 0$ in (3.8). Moreover, it makes sense that the density $\rho(0;t)$ is null when $t < t_*$ since all the poles β_n are located above the inviscid branch points x_s on the imaginary axis, and $|x_s| > 0$.

3.1. Residue computation of the integral representation of the inviscid limit for $t = t_*$.

Since $x_s(t_*) = 0$, we have

$$\int_0^\infty \frac{z^{1/3}}{x^2+z^2}\,dz = \frac{\pi}{\sqrt{3}\,x^{2/3}}. \tag{3.10}$$

Combining (3.4) and (3.10), we recover the inviscid solution at $t = t_*$:

$$u(x,t_*) = \lim_{\nu\to 0^+} u_\nu(x,t_*) = \frac{x}{t_*} - \left(\frac{x}{4t_*^4}\right)^{1/3} = \frac{x}{t_*} - \frac{U(x,t_*)}{t_*}. \tag{3.11}$$

For $0 < t < t_*$, a similar computation can be done using the double keyhole contour of integration displayed in Fig. 3.1.

For $t \geq t_*$, there is an interesting special case: since $z_s(0;t) \in \mathbb{R}$ for $t \geq t_*$, $\rho(0;t) = z_s(0;t)/\pi t$. Thus the inviscid solution at the origin (shock) for $t > t_*$ is given by (see (4.4b))

$$u(0,t) = \frac{u(0^-,t) - u(0^+,t)}{2} = \pi\rho(0;t) = \frac{1}{2}\,(t-t_*)^{1/2}t^{-3/2}\,t_*^{-1/2}, \tag{3.12}$$

i.e., the solution at the shock satisfies the jump condition (see (C.6), (4.4b), Fig. C.1, and [3, 16]):

$$u(0^{\mp},t) = \pm\frac{1}{2}\,(t-t_*)^{1/2}t^{-3/2}t_*^{-1/2}. \tag{3.13}$$

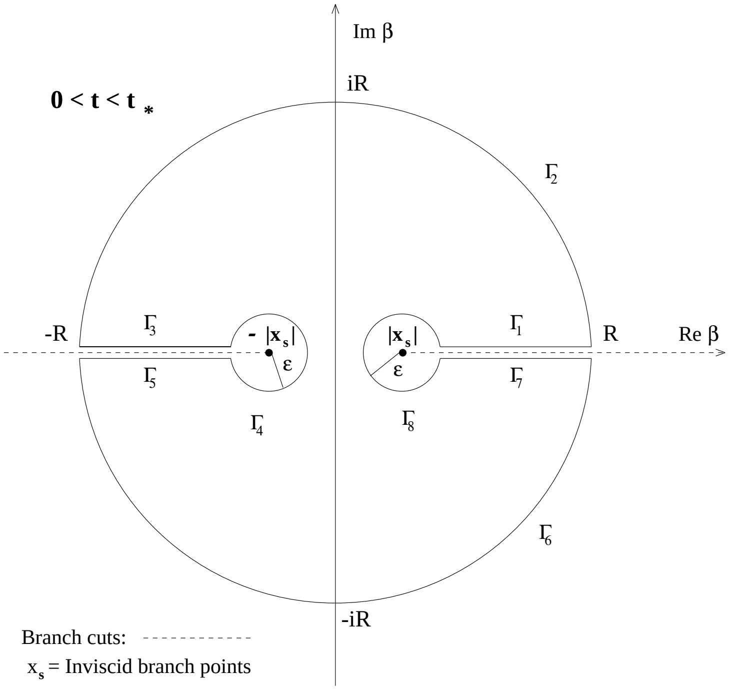


FIG. 3.1. *Contour of integration for the inviscid limit for $t < t_*$.*

Although this result was stated by Bessis and Fournier and can be derived by taking $\lim_{\beta\to 0} \rho(\beta;t)$, it is mentioned here to verify the validity formula $\rho(\beta;t) = (\pi t)^{-1}\,\Re z_s^+(\beta;t)$, where $z_s^+(\beta;t)$ is the saddle point relevant to the asymptotic expansion with a positive real part.

As a final remark, we would like to point out that this procedure which consists in recovering the analytic structure of the inviscid solution via the limiting pole density and the pole expansion is no longer possible when $t > t_*$ and $x \neq 0$. Indeed, in this case we are faced with the same (apparent) paradox that is present in the asymptotic expansion of the (spatial) Fourier transform of the inviscid solution (see section 6 and also [3]). Thus the only way to recover the inviscid solution for $t > t_*$ and $x \neq 0$ using the limiting pole density is by extending the solution obtained for $t < t_*$ to $t > t_*$.

4. Analytic extension of the integral representation of the inviscid solution on the imaginary axis. Let

$$u(x,t) = \frac{x}{t} - 2x\int_0^\infty \frac{\rho(\beta;t)}{x^2+\beta^2}\,d\beta = \frac{x}{t} - \int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{x-i\beta}\,d\beta, \tag{4.1a}$$

$$\tilde{u}(y,t) = \frac{y}{t} + 2y\,P.V.\int_0^\infty \frac{\rho(\beta;t)}{y^2-\beta^2}\,d\beta = \frac{y}{t} + P.V.\int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{y-\beta}\,d\beta. \tag{4.1b}$$

Then one can show that $\rho(y;t)$ is a density function which satisfies the conservation equation

$$\rho_t + (\rho\,\tilde{u})_y = 0. \tag{4.2}$$

Indeed, one only needs to verify that $u(x,t)$ defined by (4.1a) satisfies the inviscid BE $u_t + uu_x = 0$ under the assumption that (4.2) holds. Since u has branch cuts on

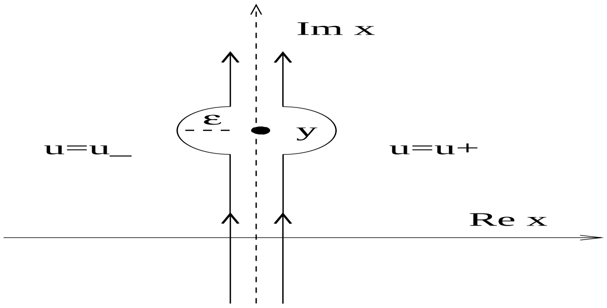


FIG. 4.1. *Analytic continuation of the integral representation of the inviscid solution on the imaginary axis.*

the imaginary axis for $t < t_*$, $\tilde{u}$ has branch cuts for y real (i.e., also on the imaginary axis); one can analytically continue u on the imaginary axis using the paths displayed in Fig. 4.1. Define the solution on the left (u_-) and right (u_+) of the imaginary axis by

$$u_\pm(iy,t) = \lim_{x\to 0^\pm} u(z = x+iy,t). \tag{4.3}$$

The discontinuity at $x = 0$ characterizes the shock solution. Thus we find that

$$\begin{aligned} u_+(iy,t) &= \frac{iy}{t} - P.V.\int_{-\infty}^{\infty} \frac{\rho(\beta;t)}{iy-i\beta}\,d\beta - \frac{\rho(y;t)}{i}\overbrace{\left(-\frac{1}{2}\oint_{|y-\beta|=\epsilon}\frac{d\beta}{y-\beta}\right)}^{\pi i} \\ &= i\left\{\frac{y}{t} + \int_{-\infty}^{\infty}\frac{\rho(\beta;t)}{y-\beta}\,d\beta + i\pi\rho(y;t)\right\}. \end{aligned}$$

Similarly, we have that

$$u_-(iy,t) = i\left\{\frac{y}{t} + \int_{-\infty}^{\infty}\frac{\rho(\beta;t)}{y-\beta}\,d\beta - i\pi\rho(y;t)\right\}.$$

Therefore,

$$\tilde{u}(y,t) = \frac{1}{2i}(u_+(iy,t) + u_-(iy,t)), \tag{4.4a}$$

$$\rho(y;t) = \frac{1}{2\pi}(u_-(iy,t) - u_+(iy,t)). \tag{4.4b}$$

Since $u_\pm$ are real on the real axis, it is clear that they satisfy the symmetry relations

$$\begin{aligned} u_\pm(\overline{x},t) &= \overline{u_\pm(x,t)}, \\ u_-(-x,t) &= -u_+(x,t), \end{aligned}$$

and therefore

$$u_\pm(iy,t) = \overline{u_\pm(-iy,t)} = -\overline{u_\mp(iy,t)}.$$

From this we have

$$\tilde{u}(y,t) = \Im u_+(iy,t), \tag{4.5a}$$

$$\rho(y;t) = -\frac{1}{\pi}\Re u_+(iy,t) = \frac{1}{\pi}\Re u_-(iy,t), \tag{4.5b}$$

and the symmetry relations

$$\tilde{u}(-y,t) = -\tilde{u}(y,t),$$
$$\rho(-y;t) = \rho(y;t).$$

Thus we have the following property.

PROPERTY 4.1.

$$u_{\pm}(iy,t) = \mp\pi\rho(y;t) + i\,\tilde{u}(y,t). \tag{4.6}$$

5. Continuum limit of the pole expansion and the CDS. Let $a_n(t,\nu) = i\beta_n(t,\nu)$, and define the complex map $\mathcal{F}(\zeta,\nu,t)$ as

$$a_n(t,\nu) = \mathcal{F}(\zeta_n^\nu = \nu n, \nu, t) : \mathbb{Z}^* \times \mathbb{R}_+^2 \to i\,\mathbb{R}_+, \quad a_{-n} = -a_n. \tag{5.1}$$

Reformulating the Mittag–Leffler expansion of u_ν (1.5) and the CDS (1.6) in terms of a_n, we have

$$u_\nu(x,t) = \frac{x}{t} - 2\nu \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{1}{x-a_l}, \tag{5.2a}$$

$$\dot{a}_n = \frac{a_n}{t} - 2\nu \sum_{\substack{l=-\infty \\ l\neq n,0}}^{\infty} \frac{1}{a_n - a_l} \quad \forall n \in \mathbb{Z}^*. \tag{5.2b}$$

Both symmetric sums (5.2a, b) should be understood as

$$\sum_{\substack{l=-\infty \\ l\neq n,0}}^{\infty} \frac{1}{a_n - a_l} = \frac{1}{2a_n} + 2a_n \sum_{\substack{l=1 \\ l\neq n}}^{\infty} \frac{1}{a_n^2 - a_l^2} = \frac{1}{2a_n} + a_n \sum_{\substack{l=-\infty \\ l\neq 0,\pm n}}^{\infty} \frac{1}{a_n^2 - a_l^2}$$

and

$$\sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{1}{x-a_l} = 2x \sum_{l=1}^{\infty} \frac{1}{x^2 - a_l^2} = x \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{1}{x^2 - a_l^2}.$$

At t_*, we have (cf. [18, section 4.1])

$$\begin{aligned} a_n(t_*,\nu) &= \mathcal{F}(\zeta_n^\nu = \nu n, \nu, t_*) = i \cdot 4t_*\,(2\nu\mu_n)^{3/4} \\ &= i \cdot 4t_*\,(2\nu(c_{-1}n + c_0 + c_1/n + \cdots))^{3/4} \\ &= i \cdot 4t_*\,\left(c_{-1}(2\nu n) + c_0 2\nu + c_1(2\nu)^2/(2\nu n) + \cdots)\right)^{3/4}. \end{aligned}$$

Introduce the map

$$f(\zeta,t) = \mathcal{F}(\zeta,0,t) : \mathbb{R}\times\mathbb{R}_+ \to i\,\mathbb{R}_+, \qquad f(-\zeta,t) = -f(\zeta,t), \tag{5.3}$$

where the continuous variable ζ corresponds to a position on the real axis which can be thought of as a variable obtained by simultaneously letting $\nu \to 0^+$ and $n \to +\infty$. Assume that

$$a_n(t,\nu) = \mathcal{F}(n\nu,\nu,t) = f(n\nu,t) + e_n(\nu,t), \tag{5.4}$$

in which $e_n(\nu, t)$ is a small error term that goes to 0 as $\nu \to 0^+$. Thus, formally we have

$$\begin{aligned} 2\nu \sum_{\ell \neq n} \frac{1}{a_n(t,\nu) - a_\ell(t,\nu)} &\simeq 2\nu \sum_{\ell \neq n} \frac{1}{f(n\nu, t) - f(\ell\nu, t)} \\ &\xrightarrow{\nu \to 0^+} P.V. \int_{-\infty}^{\infty} \frac{d\zeta'}{f(\zeta, t) - f(\zeta', t)}. \end{aligned} \tag{5.5}$$

Moreover, this approximation shows that representation (5.4) is valid for all time if it is true at $t = t_*$. A rigorous analysis of approximation (5.5) has been performed in the context of vortex sheets in [10]. It is then clear that the pair of equations (5.2a, b) satisfy the following property.

PROPERTY 5.1. *The continuum limit of the CDS and the pole expansion is the parametric system of integro-differential equations defined for any x such that $\forall \zeta \in \mathbb{R}$, $x \neq f(\zeta, t)$, by*

$$\begin{aligned} u(x,t) &= \frac{x}{t} - \int_{-\infty}^{\infty} \frac{d\zeta'}{x - f(\zeta', t)}, \\ \frac{\partial f}{\partial t}(\zeta, t) &= \frac{f(\zeta, t)}{t} - P.V. \int_{-\infty}^{\infty} \frac{d\zeta'}{f(\zeta, t) - f(\zeta', t)}. \end{aligned}$$

This property also can be expressed as

$$\frac{\partial f}{\partial t}(\zeta, t) = \frac{f(\zeta, t)}{t} - f(\zeta, t)\, P.V. \int_{-\infty}^{\infty} \frac{d\zeta'}{f^2(\zeta, t) - f^2(\zeta', t)} \tag{5.6}$$

and

$$u(x,t) = \frac{x}{t} - x \int_{-\infty}^{\infty} \frac{d\zeta'}{x^2 - f^2(\zeta', t)}, \qquad x \neq f(\zeta, t). \tag{5.7}$$

Equation (5.7) defines the branch cuts of the inviscid solution as the set of complex x-points for which $x = f(\zeta, t)$, while equation (5.6) defines the dynamics of these branch cuts.

Since the poles are located on the imaginary axis, one can make the additional assumption that $f(\zeta, t) = i\, g(\zeta, t)$, where $g(\zeta, t) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$. Then from (5.7) we find

$$u(x,t) = \frac{x}{t} - 2x \int_{0}^{\infty} \frac{d\zeta'}{x^2 + g^2(\zeta', t)}, \qquad x \neq i\, g(\zeta, t). \tag{5.8}$$

We then define the density function $\rho(z, t)$ as

$$\rho(z, t) = \frac{1}{g_\zeta(\zeta, t)}, \tag{5.9}$$

where $z = g(\zeta, t)$. Then $d\zeta' = \rho(z', t)\, dz'$ and we introduce this change of variable in (5.8) to obtain

$$u(x,t) = \frac{x}{t} - 2x \int_{0}^{\infty} \frac{\rho(z; t)}{x^2 + z^2}\, dz.$$

As a converse to the procedure of section 3, case (i) $t = t_*$, the pole positions at t_* can be recovered from the cumulative distribution function σ (see Definition 2.1):

$$\sigma(\beta; \beta_{\max}, t) = \int_0^{\beta} \rho(\xi; t)\, d\xi \quad \text{for } |\beta| \leq \beta_{\max} \tag{5.10}$$

and, in particular, for $t = t_*$, $\rho(\beta; t_*) = \frac{2\sqrt{3}}{\pi}(4t_*)^{-4/3}|\beta|^{1/3}$, so

$$\begin{aligned}\sigma(\beta; \beta_{\max}, t_*) &= \int_0^{\beta_*} \frac{2\sqrt{3}}{\pi}(4t_*)^{-4/3}\xi^{1/3}\, d\xi \\ &= \left(\frac{2\pi}{3\sqrt{3}}\right)^{-1} \left(\frac{\beta_*}{4t_*}\right)^{4/3}.\end{aligned}$$

Inverting this relation in terms of β_*, we find that

$$\beta_* = 4t_* \left(\frac{2\pi}{3\sqrt{3}}\sigma_*\right)^{3/4}. \tag{5.11}$$

In order to recover the correct discretization, it suffices to choose

$$\sigma_* = \sigma_n(t_*, \nu) = 2\nu\mu_n \left(\frac{2\pi}{3\sqrt{3}}\right)^{-1}, \tag{5.12}$$

where $\mu_n = \frac{2\pi}{3\sqrt{3}}(n - \frac{1}{3}) + \mathcal{O}(1/n)$ as $n \to +\infty$ (see [18, section 4.1]). Therefore,

$$\sigma_* = 2\nu(n - 1/3) + \mathcal{O}(1/n) \quad \text{as } n \to +\infty, \tag{5.13}$$

and

$$\beta_* = \beta_n(t_*, \nu) = 4t_* \left(2\nu\mu_n\right)^{3/4}. \tag{5.14}$$

Similar computations can be found in [21].

6. Uniform asymptotic expansion in $(0, t_*]$ of the spatial Fourier transform of the inviscid solution $\hat{u}(k, t)$ as $k \to +\infty$.

The analyticity properties of the inviscid solution also can be analyzed by describing the asymptotic behavior of its Fourier transform (see [13, 20]). We find a uniform asymptotic expansion as $k \to +\infty$ of the Fourier transform of the inviscid solution in a neighborhood of $t = t_*$, where two second-order branch points $\pm x_s(t)$ coalesce into a third-order branch point at the origin $x_s(t_*) = 0$. Thus, we clarify the seemingly discontinuous change of behavior of $\hat{u}(k, t)$ at t_* presented in [13]. This result is resumed in the following theorem.

THEOREM 6.1. *The uniform asymptotic expansion of the Fourier transform of the inviscid solution for* $0 < t \leq t_*$ *is*

$$\hat{u}(k, t) = \mathcal{C}_0 \cdot (tk)^{-4/3} Ai\left[(-3ikx_s(t)/2)^{2/3}\right]\left(1 + \mathcal{O}\left(k^{-1}\right)\right) \quad \textit{as } k \to +\infty.$$

Thus from the asymptotic property of the Airy function and its value at the origin $Ai(0)$ *we have*

$$\hat{u}(k, t) \sim \begin{cases} \mathcal{C}_1(t) \cdot (t_* - t)^{-1/4}\, k^{-3/2}\, e^{-k\,|x_s(t)|} & 0 < t < t_* \\ \mathcal{C}_2 \cdot (t_* k)^{-4/3} & t = t_* \end{cases} \quad \textit{as } k \to +\infty.$$

Proof. In [13], Fournier and Frisch derive the asymptotic behavior of the inviscid solution via the so-called Fourier–Lagrangian (F–L) representation, which is valid up to the time where the relation $x(x_0,t) = x_0 + tu_0(x_0)$ is invertible, i.e., up to $t_* = -(\inf_{x_0} u_0'(x_0))^{-1}$. There is a discontinuous change in the behavior of $\hat{u}(k,t)$ in $k^{-3/2}\exp(-k|x_s(t)|)$ before t_* to $(t_*k)^{-4/3}$ at t_*, which arises from the fact that the two saddle points of multiplicity 1 for $0 < t < t_*$ coalesce at the origin to form a saddle point of multiplicity 2 at $t = t_*$. The F–L representation is found by changing variables from the Eulerian coordinate to the Lagrangian coordinate, followed by an integration by parts:

$$\begin{aligned}\hat{u}(k,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-ikx}u(x,t)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-ikx(x_0,t)}u_0(x_0)\frac{\partial x}{\partial x_0}dx_0\\ &= \frac{1}{\sqrt{2\pi}ik}\int_{-\infty}^{\infty} e^{-ikx(x_0,t)}u_0'(x_0)\,dx_0 \qquad (k\neq 0).\end{aligned}$$

For $u_0(x) = 4x^3 - x/t_*$, $x(x_0,t) = 4tx_0^3 + x_0(1-t/t_*)$, we find

$$\hat{u}(k,t) = \frac{1}{\sqrt{2\pi}ik}\int_{-\infty}^{\infty}\exp\left\{-12ikt\left(\frac{x_0^3}{3} - \frac{\alpha}{6}\,x_0\right)\right\}u_0'(x_0)\,dx_0,$$

where $2\alpha = 1/t_* - 1/t$. Let $\lambda = -12ikt$; then we are interested in finding the behavior of $\hat{u}(k,t)$ as $k\to+\infty$, that is, the behavior as $\lambda\to\infty$ of the integral

$$\int_{-\infty}^{\infty}\exp\left\{\lambda\left(\frac{x_0^3}{3} - \frac{\alpha}{6}\,x_0\right)\right\}u_0'(x_0)\,dx_0. \tag{6.1}$$

The saddle points of the integrand occur when $\partial x/\partial x_0 = 0$; thus

$$x_0 = x_0^{\pm}(t) = \pm\sqrt{\frac{\alpha}{6}} = \pm\sqrt[3]{\frac{x_s(t)}{8t}} \Rightarrow x\left(x_0^{\pm}(t)\right) = \pm x_s(t).$$

At $t = t_*$, $x_0^{\pm}(t_*) = 0$, and the two saddle points of multiplicity 1 have coalesced into a saddle point of multiplicity 2 at the origin. Let

$$f(x_0) = \frac{x(x_0,t)}{12t} = \frac{x_0^3}{3} - \frac{\alpha}{6}\,x_0,$$

and recall that $x_s(t) = t(2\alpha/3)^{3/2}$ (see (C.4)). Then

$$f(x_0^{\pm}(t)) = -\frac{2}{3}x_0^{\pm}(t)^3 = \mp\frac{x_s(t)}{12t}.$$

We introduce the coefficients

$$\begin{aligned}\zeta^{3/2} &= \frac{3}{4}\left(f(x_0^-(t)) - f(x_0^+(t))\right) = \frac{3}{2}f(x_0^-(t)) = \frac{x_s(t)}{8t},\\ \eta &= \frac{1}{2}\left(f(x_0^-(t)) + f(x_0^+(t))\right) = 0,\end{aligned}$$

which arise in the construction of a uniform asymptotic expansion of an integral with two coalescing saddle points. The integral defined in (6.1) is already in a format appropriate for such a derivation. Indeed, it is an integral of the form

$$I(\lambda;\zeta,\eta) = \int_{\mathcal{C}}\exp\left\{\lambda\left(u^3/3 - \zeta u + \eta\right)\right\}\phi_0(u)\,du,$$

where $\lambda \to \infty$. Thus the 1–1 analytic transformation $x_0 \leftrightarrow u$ given by the equation $f(x_0) = u^3/3 - \zeta u + \eta$ is simply the identity $x_0 \equiv u$. Therefore, the time-uniform asymptotic expansion of the spatial Fourier transform of the solution $\hat{u}(k,t)$ is immediately found in terms of the Airy function and its derivative (cf. [11, 23, section VII-4]). Let

$$a_0 = \frac{1}{2}\left[u_0'(\zeta^{1/2}) + u_0'(-\zeta^{1/2})\right], \qquad b_0 = \frac{1}{2\zeta^{1/2}}\left[u_0'(\zeta^{1/2}) - u_0'(-\zeta^{1/2})\right],$$

then

$$\hat{u}(k,t) = \frac{e^{-\lambda\eta}}{\sqrt{2\pi}ik} \cdot 2\pi i \left[\frac{\mathrm{Ai}\left[\lambda^{2/3}\zeta\right]}{\lambda^{1/3}}(a_0 + \mathcal{O}(1/\lambda)) + \frac{\mathrm{Ai}'\left[\lambda^{2/3}\zeta\right]}{\lambda^{2/3}}(b_0 + \mathcal{O}(1/\lambda))\right]$$

as $\lambda(k) = -12ikt \to \infty$. Since $\eta = 0$, $\zeta^{3/2} = x_s(t)/8t$, $\zeta^{1/2} = x_0^+(t)$, $u_0'(x_0) = 12x_0^2 - 1/t_*$, $a_0 = u_0'(x_0^\pm(t)) = 1/t$, $b_0 = 0$, and we obtain the asymptotic behavior of $\hat{u}(k,t)$ as $k \to +\infty$ uniform in a compact interval containing $t = t_*$:

$$\hat{u}(k,t) = \mathcal{C}_0 \cdot (tk)^{-4/3}\mathrm{Ai}\left[(-3ikx_s(t)/2)^{2/3}\right]\left(1 + \mathcal{O}\left(k^{-1}\right)\right) \quad \text{as } k \to +\infty, \tag{6.2}$$

where $\mathcal{C}_0$ is an appropriate numerical constant. Note that (6.2) can be obtained without recourse to this method by a classical asymptotic analysis in which one would express the expansion in terms of the Airy function. We choose this derivation due to the simplicity of its construction. For $0 < t < t_*$, using the fact that

$$\mathrm{Ai}(z) = \frac{e^{-\frac{2}{3}z^{3/2}}}{2\sqrt{\pi}z^{1/4}}\left(1 + \mathcal{O}\left(z^{-3/2}\right)\right) \quad \text{as } z \to \infty \text{ in } |\arg z| < \pi$$

and for $t = t_*$, and by evaluating $\mathrm{Ai}(0) = 3^{-2/3}/\Gamma(2/3)$, we obtain the asymptotic behavior of $\hat{u}(k,t)$ as $k \to +\infty$ for $0 < t \le t_*$ described in the second part of Theorem 6.1, in which $\mathcal{C}_1(t)$ is a constant depending on t and $\mathcal{C}_2$ is a numerical constant. These expansions are consistent with the fact that the Fourier transform of an analytic function with a branch-point singularity at $x_s = \Re x_s + i\Im y_s$ of the form (cf. [20])

$$v(z) \sim (z - x_s)^\mu, \qquad \mu \notin \mathbb{Z},$$

has an asymptotic behavior of the form

$$\hat{v}(k) \sim k^{-(\mu+1)}e^{-k\Im x_s}e^{ik\Re x_s} \quad \text{as } k \to +\infty.$$

Note that the expansion for $t > t_*$ obtained from (6.2) yields the incorrect behavior $|\hat{u}(k,t)| \sim \mathcal{C}_1(t) \cdot (t - t_*)^{-1/4}\, k^{-3/2}$, which is valid only for moderate wave numbers of the form $1 \ll k \le 1/|x_s|$. This is due to the fact that the formal F–L representation is no longer valid beyond t_*. The correct behavior after t_* of the form $|\hat{u}_I(k,t)| \sim \mathcal{C}_3(t) \cdot (t - t_*)^{1/2}k^{-1}$ for $k > 1/|x_s|$ which reflects the presence of a shock then must be obtained by following the work of Fournier and Frisch in [13]. The two expansions agree when $k \simeq 1/|x_s|$, giving a behavior of the form $|\hat{u}(k,t)|, |\hat{u}_I(k,t)| \sim \mathcal{C}_4(t) \cdot (t - t_*)^2$ for t close to t_* (see [13]).

Appendix A. On the generic nature of the initial data. Caflisch et al. characterize geometrically generic singularities for nonlinear hyperbolic systems in [9] in the following way: given a PDE and its initial data, a singularity is generic if, under perturbation of the "initial data," the singularity is of one of the stable types, namely, a fold corresponding to a square-root branch point in z for each t or a cusp corresponding to a cube root branch point which occurs when the two square root branch points collide. They show that these are the only stable singularity types for the inviscid BE. Loosely, they define stability as the property that under perturbation of the initial data, the perturbed solution will have the same singularity type as the original problem, i.e., either a fold or a cusp. Note that the formation of a cube root singularity must stem from a "tangential" collision of the square root branch points, i.e., one where the branch points travel at the same characteristic speed. In case of a "nontangential" collision of square root branch points travelling at different characteristic speeds, the resulting singularity remains a square root branch point. For more details see [9].

Fournier and Frisch characterized generic singularities and corresponding generic initial data for the inviscid BE in [13]. This description is based on a local analysis of the singularity and takes into account the Gallilean invariance of the PDE and its invariance under translation of the reference frame. This was reformulated in Bessis and Fournier's first paper [4].

Appendix B. Cardan's formula. The roots of a cubic polynomial are given by the well-known formula of Cardan (cf. [1, section 3.8.2]). We state this formula to clarify the choices that are made in choosing the branches of the algebraic functions which define the saddle points in the expansions: let λ, a, b, $c \in \mathbb{C}$; then the roots of the equation

$$\lambda^3 + a\lambda^2 + b\lambda + c = 0 \tag{B.1}$$

are obtained by setting

$$A = a/3, \quad B = b/3, \quad \alpha = A^2 - B, \quad \zeta = 2A^3 - 3AB + c.$$

Let $\lambda = x - A$. Then (B.1) becomes

$$x^3 - 3\alpha x + \zeta = 0. \tag{B.2}$$

Let $\omega = e^{2\pi i/3}$ be a cube root of unity; then the three roots of (B.2) are

$$\left\{\begin{array}{l} x_0 = \omega\mathcal{A} + \omega^2\mathcal{B}, \\ x_1 = \omega^2\mathcal{A} + \omega\mathcal{B}, \\ x_2 = \mathcal{A} + \mathcal{B}, \end{array}\right. \qquad \text{where} \qquad \left\{\begin{array}{l} \Delta = (\zeta/2)^2 - \alpha^3, \\ \mathcal{A} = \sqrt[3]{-\zeta/2 + \sqrt{\Delta}}, \\ \mathcal{B} = \sqrt[3]{-\zeta/2 - \sqrt{\Delta}}. \end{array}\right. \tag{B.3}$$

After choosing a branch for $\mathcal{A}$, one must choose the corresponding branch for $\mathcal{B}$ so that $\mathcal{A} \cdot \mathcal{B} = \alpha^3$. If α and ζ are real, then there are three possibilities depending on the sign of the real discriminant Δ:

(i) $\Delta < 0: \ \mathcal{A}, \mathcal{B} \in \mathbb{C}, \ \mathcal{A} = \overline{\mathcal{B}}, x_0, x_1, x_2 \in \mathbb{R}$;
(ii) $\Delta = 0: \ \mathcal{A} = \mathcal{B} \in \mathbb{R}, \ x_0 = x_1 = -\frac{x_2}{2} \in \mathbb{R}$;
(iii) $\Delta > 0: \ \mathcal{A}, \mathcal{B} \in \mathbb{R}, x_0 = \overline{x_1} \in \mathbb{C}, \ x_2 \in \mathbb{R}$.

Case (iii), which yields two conjugate roots, is the only instance when we can expect to have two equally relevant saddle points, thus allowing for some cancellation in the asymptotic expansion. The relevant roots x_0 and x_1, after separation of real and imaginary parts, are given by

$$x_0 = \overline{x_1} = -\frac{1}{2}(\mathcal{A} + \mathcal{B}) + i\frac{\sqrt{3}}{2}(\mathcal{A} - \mathcal{B}). \tag{B.4}$$

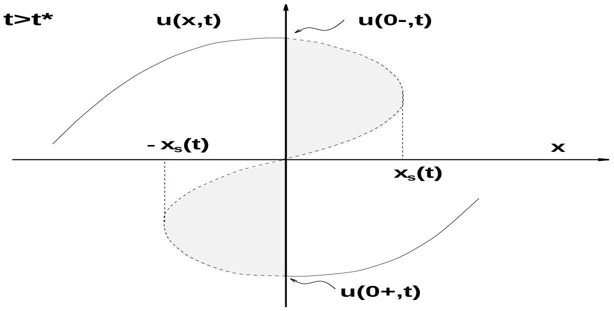


FIG. C.1. *Shock, multivaluedness, branch points, and Maxwell's equal area rule for* $t > t_*$.

Appendix C. Inviscid solution ($\nu = 0$). The inviscid BE states that the velocity of a fluid particle is conserved along certain trajectories, namely, the characteristic lines

$$\dot{x} = \frac{dx}{dt}(t) = u(x(t), t) \tag{C.1}$$

in the (x, t) plane. The implicit solution obtained by the method of characteristics reflects the conservation of the velocity along these special curves:

$$\begin{cases} u = u(x,t) = u_0(x_0(x,t)), \\ x = x_0 + t\, u_0(x_0(x,t)). \end{cases} \tag{C.2}$$

A fluid particle originally at a (Lagrangian) position x_0 in space will be at a new (Eulerian) position x after a certain time t with the same velocity along this line. Let $U = x_0$. Then, substituting $u_0(x) = 4x^3 - x/t_*$ in (C.2), we find that U satisfies the cubic equation

$$U^3 - \frac{\alpha}{2}U - \frac{x}{4t} = 0, \qquad \alpha = \frac{t - t_*}{2tt_*}. \tag{C.3}$$

This defines a three-sheeted Riemann surface for the solution with a third-order branch point at infinity and two opposite second-order branch points at $\pm x_s(t)$ defined by

$$x_s(t) = t\,(2\alpha/3)^{3/2} = i\,(3t_*)^{-3/2}(t_* - t)^{3/2}t^{-1/2}. \tag{C.4}$$

The envelope of the characteristic lines is the branch point since $0 = \frac{\partial x}{\partial x_0} \Rightarrow x\left(x_0^{\pm}(t)\right) = \pm x_s(t)$. The solution is therefore

$$U(x,t) = \begin{cases} (8t)^{-1/3}\left\{\sqrt[3]{x + \sqrt{x^2 - x_s^2}} + \sqrt[3]{x - \sqrt{x^2 - x_s^2}}\right\} & t \neq t_*, \\ \sqrt[3]{\frac{x}{4t_*}} & t = t_*. \end{cases} \tag{C.5}$$

Note the particular (real) values of $u(x,t)$ at the shock at $x = 0$ already found in (3.13):

$$u(0^{\pm}, t) = -U(0^{\pm}, t)/t = \begin{cases} \mp\frac{1}{2}\,(t - t_*)^{1/2}t^{-3/2}t_*^{-1/2} & t \geq t_*, \\ 0 & t < t_*. \end{cases} \tag{C.6}$$

The topology of the three-sheeted Riemann surface given by (C.5) and the interpretation of the shock as the permutation of two Riemann sheets has been fully explained by Bessis and Fournier in [3].

Appendix D. Generalization of the initial data to $u_0(x) = 2nx^{2n-1} - x/t_*$, $n \geq 2$. Although the inviscid singularity resulting from a polynomial of arbitrary odd order of the form $u_0(x) = 2nx^{2n-1} - x/t_*$, $n \in \mathbb{N}$, $n \geq 2$, is no longer generic (see [13, p. 707]), it is still interesting to describe the behavior of the inviscid solution and the related asymptotic density of poles. We first describe the inviscid solution and its branch points: substituting $u_0(x)$ in (C.2), we find that U is a root of the polynomial of degree $2n-1$. Let

$$\alpha_n(t) = \frac{t - t_*}{ntt_*} \quad \text{and} \quad P_n(U) = U^{2n-1} - \frac{\alpha_n(t)}{2}\,U - \frac{x}{2nt} = 0.$$

Let $U_s(t)$ satisfy

$$0 = \frac{\partial x}{\partial U}(U_s(t)) = P_n'(U_s(t)) \Longrightarrow U_s(t) = \left(\frac{\alpha_n(t)/2}{2n-1}\right)^{1/(2n-2)}.$$

The $2n-2$ branch points of the inviscid solution are then given by

$$x_s(t) = x(U_s(t)) = 2ntU_s\left(U_s^{2n-2} - \alpha_n(t)/2\right) = \mathcal{C}_n \cdot (t - t_*)^{\frac{2n-1}{2n-2}} \cdot t_*^{-\frac{2n-1}{2n-2}} \cdot t^{-\frac{1}{2n-2}},$$

where $\mathcal{C}_n = -(2n-2)(2n-1)^{\frac{2n-1}{2n-2}}(2n)^{-\frac{1}{2n-2}}$. Notice that the $2n-2$ branch points coalesce at the origin at t_*. Since $U(0^{\pm}, t) = \pm\pi t\rho(0; t) = \pm(\alpha_n(t)/2)^{\frac{1}{2n-2}}$ (cf. (3.9)), the (real) value of the inviscid solution at the origin (shock) is given by

$$u(0^{\pm}, t) = -U(0^{\pm}, t)/t = \mp\pi\rho(0; t) = \begin{cases} \mp(t - t_*)^{\frac{1}{2n-2}}(2nt_*)^{\frac{-1}{2n-2}}t^{-\frac{2n-1}{2n-2}} & t \geq t_*, \\ 0 & t < t_*. \end{cases}$$

The limiting pole density and inviscid limit are obtained using results from [17], where it is shown that the relevant saddle points in the asymptotic analysis of $E_\nu(i\beta, t_*)$ are

$$z_0(\beta; t_*) = \exp\left(\frac{i\pi}{4n-2}\right)\left(\frac{\beta}{2nt_*}\right)^{\frac{1}{2n-1}}, \quad z_1 = -\overline{z_0}.$$

Since

$$\rho(\beta; t_*) = \frac{\Re z_0(\beta; t_*)}{\pi t_*}$$

and further results in [17] concern the asymptotic behavior of the zeros $\mu_{k,n}$ of $\mathcal{F}_n(\mu) = \int_{-\infty}^{\infty} e^{\mu(2niz - z^{2n})}dz$, we have the following result.

Property D.1. *For any integer $n \geq 2$, the density of poles at the shock time t_* arising from the initial data $u_0(x) = 2nx^{2n-1} - x/t_*$ is*

$$\rho(\beta; t_*) = \frac{1}{\pi}\cos\left(\frac{\pi}{4n-2}\right)\left(\frac{\beta}{2nt_*^{2n}}\right)^{\frac{1}{2n-1}}.$$

The density at the origin ($z = 0$) *is*

$$\rho(0;t) = \begin{cases} \frac{1}{\pi t}\left(\frac{t-t_*}{2ntt_*}\right)^{\frac{1}{2n-2}} & t > t_*, \\ 0 & t \leq t_*. \end{cases}$$

Moreover at t_*, *the* k*th ordered pole of the solution for* $\nu > 0$ *is located at*

$$a_{k,n}(t_*, \nu) = i \cdot 2nt_* \left(2\nu\mu_{k,n}\right)^{\frac{2n-1}{2n}},$$

where the positive coefficients $\mu_{k,n}$ *are asymptotically given by*

$$\mu_{k,n} = \frac{\pi}{4n-2}\sec\left(\frac{\pi}{4n-2}\right)\left(\frac{n-1}{2n-1} + 1 + 2k\right) + \mathcal{O}\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty.$$

Higher order approximations of $\mu_{k,n}$ are provided in [17]. At $t = t_*$, the inviscid solution can be found via the complex-valued limiting pole density $\rho(\beta; t_*)$ and the pole expansion as in (3.11) or via the characteristic equation as in Appendix C. The resulting singularity is a branch point of order $2n-1$, which arises from the coalescence of $n-1$ pair(s) of conjugate branch points of order $2n-2$ (see [13, p. 707]):

$$u(x, t_*) = \frac{x}{t_*} - \left(\frac{x}{2nt_*^{2n}}\right)^{\frac{1}{2n-1}}.$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] L. V. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.

[3] D. Bessis and J. D. Fournier, *Complex singularities and the Riemann surface for the Burgers equation*, Research Reports in Physics: Nonlinear Physics, Springer-Verlag, Berlin, New York, 1990, pp. 252–257.

[4] D. Bessis and J. D. Fournier, *Pole condensation and the Riemann surface associated with a shock in Burgers equation*, J. Phys. Lett., 45 (1984), L833–L841.

[5] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[6] J. M. Burgers, *The Nonlinear Diffusion Equation*, D. Reidel, Boston, MA, 1974.

[7] J. M. Burgers, *A mathematical model illustrating the theory of turbulence*, Adv. Appl. Mech., 1 (1948), pp. 171–199.

[8] F. Calogero, *Motion of poles and zeros of special solutions of nonlinear and linear partial differential equations and related "solvable" many body problems*, Nuovo Cimento B (11), 43 (1978), pp. 177–241.

[9] R. E. Caflisch, N. Ercolani, T. Y. Hou, and Y. Landis, *Multi-valued solutions and branch point singularities for nonlinear hyperbolic or elliptic systems*, Comm. Pure Appl. Math., 46 (1993), pp. 453–499.

[10] R. E. Caflisch and J. S. Lowengrub, *Convergence of the vortex method for vortex sheets*, SIAM J. Numer. Anal., 26 (1989), pp. 1060–1080.

[11] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599–611.

[12] N. M. Ercolani, C. D. Levermore, and T. Zhang, *The behavior of the Weyl function in the zero-dispersion KdV limit*, Comm. Math. Phys., 183 (1997), pp. 119–143.

[13] J. D. Fournier and U. Frisch, *L'équation de Burgers deterministe et statistique*, J. Mech. Theory Appl., 2 (1983), pp. 699–750.

[14] Y. Kimura, *Dynamics of complex singularities for Burgers' equation*, in Proc. NEEDS '94, V.G. Makhankov, ed., World Scientific, Singapore, 1995.

[15] P.D. Lax and C.D. Levermore, *The small dispersion limit of the Korteweg–de Vries equation*, I, II, III, Comm. Pure Appl. Math., 36 (1983), pp. 253–290, pp. 571–593, pp. 809–829.

[16] T. D. Lee and C. N. Yang, *Statistical theory of equations of state and phase transitions.* II *Lattice gas and ising model*, Phys. Rev., 87 (1952), pp. 410–447.

[17] D. Senouf, *Asymptotic and numerical approximations of the zeros of Fourier integrals*, SIAM J. Math. Anal., 27 (1996), pp. 1102–1128.

[18] D. Senouf, *Dynamics and condensation of complex singularities for Burgers' equation* I, SIAM J. Math. Anal., 28 (1997), pp. 1457–1489.

[19] D. Senouf, R. Caflisch, and N. Ercolani, *Pole dynamics and oscillations for complex Burgers equation in the small dispersion limit*, Nonlinearity, 9 (1996), pp. 1671–1702.

[20] C. Sulem, P. L. Sulem, and H. Frisch, *Tracing complex singularities with spectral methods*, J. Comp. Phys., 50 (1983), pp. 138–161.

[21] O. Thual, U. Frisch, and M. Hénon, *Application of pole decomposition to an equation governing the dynamics of wrinkled flame fronts*, J. Phys., 46 (1985), pp. 1485–1494.

[22] F. Ursell, *Integrals with a large parameter. Several nearly coincident saddle points*, Proc. Cambridge Philos. Soc., 72 (1972), pp. 49–65.

[23] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.